The Vector Poisson Channel: On the Linearity of the Conditional Mean Estimator

Alex Dytso, *Member, IEEE*, Michael Fauß, *Member, IEEE*, and H. Vincent Poor, *Life Fellow, IEEE*

***Abstract*—This work studies properties of the conditional mean estimator in vector Poisson noise. The main emphasis is to study conditions on prior distributions that induce linearity of the conditional mean estimator. The paper consists of two main results. The first result shows that the only distribution that induces the linearity of the conditional mean estimator is a product gamma distribution. Moreover, it is shown that the conditional mean estimator cannot be linear when the dark current parameter of the Poisson noise is non-zero. The second result produces a quantitative refinement of the first result. Specifically, it is shown that if the conditional mean estimator is close to linear in a mean squared error sense, then the prior distribution must be close to a product gamma distribution in terms of their Laplace transforms. Finally, the results are compared to their Gaussian counterparts.**

***Index Terms*—Conditional mean estimator, conjugate priors, estimation theory, gamma distribution, gaussian noise, vector poisson noise.**

I. Introduction

THIS work considers a problem of estimating a random vector $\mathbf{X}$ from a noisy observation $\mathbf{Y}$ where $\mathbf{Y}$ given $\mathbf{X} = \boldsymbol{x}$ (denoted by $\mathbf{Y}|\mathbf{X} = \boldsymbol{x}$) follows a *vector Poisson distribution*. The objective is to characterize conditions under which the *conditional mean estimator* (i.e., $\mathbb{E}[\mathbf{X}|\mathbf{Y}]$) is a linear estimator. Conditional mean estimators are an important class of estimators that are optimal under a large family loss functions, namely Bregman divergences [1]. For example, we are interested in characterizing the set of prior distributions on $\mathbf{X}$ that induce linearity of $\mathbb{E}[\mathbf{X}|\mathbf{Y}]$. Also, we are interested in which linear estimators are realizable from $\mathbb{E}[\mathbf{X}|\mathbf{Y}]$. That is, given that $\mathbb{E}[\mathbf{X}|\mathbf{Y}] = \mathsf{H}\mathbf{Y} + \boldsymbol{c}$, what values of a matrix H and vector $\boldsymbol{c}$ are permitted? Finally, we are interested in the question of the stability of linear estimators. In other words, suppose that $\mathbb{E}[\mathbf{X}|\mathbf{Y}]$ is 'close' to a linear function, can we make statements about the distribution of $\mathbf{X}$?

Our interest in these questions stems from the wide applicability of the Poisson noise model. Indeed, the literature on the Poisson distribution is considerable, and the interested reader is referred to [2], [3] and [4] for applications of the Poisson model in compressed sensing; [5] and [6] for a summary of communication theoretic applications; [7], [8] and [9] for applications in information theory; and [10], [11] and [12] for applications of the Poisson distributions in signal processing and other fields.

Despite the wide use of the Poisson noise model in statistical science, the aforementioned questions have not been fully addressed in the vector Poisson case. The aim of this work is to fill this gap. It is interesting to note that such questions have been answered for the Gaussian noise model and are part of standard tools of statistical signal processing.

The linearity of the conditional expectation is intimately connected with the notation of *conjugate priors*, which is an important element of Bayesian statistics. In its original definition in [13, Ch. 3], the family of prior distributions is said to be conjugate if it is closed under sampling – the prior is said to be closed under sampling when both prior and posterior belong to the same family of distributions. In other words, the distribution of $\mathbf{X}$ and the distribution of $\mathbf{X}|\mathbf{Y} = \boldsymbol{y}$ are in the same family.

The structure of the conjugate prior is highly dependent on the nature of the distribution of $\mathbf{Y}|\mathbf{X} = \boldsymbol{x}$ (often termed likelihood distribution or noise distribution). For example, in [14], the authors have made considerable progress in characterizing conjugate priors for the case when the likelihood distribution belongs to the *exponential family*. In particular, in [14], it has been shown that a subset of the exponential family, characterized by certain regularity conditions, has a corresponding set of conjugate priors. Moreover, this set of conjugate priors is completely characterized by the linearity of the posterior expectation:

$$\mathbb{E}[\mathbf{X}|\mathbf{Y}] = \mathsf{H}\mathbf{Y} + \mathbf{b}, \tag{1}$$

where $\mathsf{H} = a\mathsf{I}$ for some constant a and $\mathbf{b}$ is some constant vector.

We note that the case when H is a general matrix was not considered in [14]. Moreover, even the case when $\mathbf{Y}|\mathbf{X} = \boldsymbol{x}$ follows a Poisson distribution is not covered by the regularity conditions found in [14]. However, it was shown earlier in [15] that the conjugate prior for the scalar Poisson distribution is a gamma distribution, and that the linearity of the posterior expectation holds and is a characterizing property. The proof in [15] was generalized in [16] to include several families of discrete distributions not covered by the regularity conditions of [14]. This work considers an arbitrary matrix H and characterizes

Manuscript received March 19, 2020; revised July 4, 2020; accepted September 3, 2020. Date of publication September 21, 2020; date of current version October 22, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Monica F. Bugallo. This work was supported by the U.S. National Science Foundation under Grant CCF-1908308. The work of Michael Fauß was supported by the German Research Foundation (DFG) under Grant 424522268. *(Corresponding author: Alex Dytso.)*

Alex Dytso is with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: alex.dytso@njit.edu).

Michael Fauß and H. Vincent Poor are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: mfauss@princeton.edu; poor@princeton.edu).

Digital Object Identifier 10.1109/TSP.2020.3025525

the sufficient and necessary conditions for the existence of the conjugate prior.

The paper is organized as follows. Section II presents the Poisson noise model. Section III presents and discusses our main results, which are described in Theorem 1 and Theorem 2. Section IV and Section V are dedicated to the proofs of Theorem 1 and Theorem 2, respectively. Finally, Section VI concludes the paper and discusses implications of our results by reflecting on the following: a practically relevant parametrization of a Poisson noise model, which, for example, explicitly incorporates the dark current parameter; and Gaussian noise counterparts of our results.

Notation: Throughout the paper we adopt the following notation. $\mathbb{R}^n$ denotes the space of all n-dimensional vectors, $\mathbb{R}^n_+$ the space of all n-dimensional vectors with non-negative components, and $\mathbb{Z}^n_+$ the n-dimension non-negative integer lattice. Vectors are denoted by bold lowercase letters, random vectors by bold uppercase letters, and matrices by bold uppercase sans serif letters (e.g., $\boldsymbol{x}, \mathbf{X}, \mathsf{X}$). All vectors are assumed to be column vectors. For $\boldsymbol{x} \in \mathbb{R}^n$, $\mathsf{diag}(\boldsymbol{x}) \in \mathbb{R}^{n\times n}$ denotes the diagonal matrix with the main diagonal given by $\boldsymbol{x}$. The vector with one at position i and zero otherwise is denote by $\mathbf{1}_i$. In this paper, the gamma distribution has a probability density function (pdf) given by

$$f(x) = \frac{\alpha^\theta}{\Gamma(\theta)} x^{\theta-1} \mathrm{e}^{-\alpha x}, \; x \geq 0, \tag{2}$$

where $\theta > 0$ is the shape parameter and $\alpha > 0$ is the rate parameter. We denote the distribution with the pdf in (2) by $\mathsf{Gam}(\alpha, \theta)$. Finally, the Laplace transform of the distribution of a random vector $\mathbf{U} \in \mathbb{R}^n$ is denoted by

$$\mathcal{L}_{\mathbf{U}}(\boldsymbol{t}) = \mathbb{E}\left[\mathrm{e}^{-\boldsymbol{t}^{\mathsf{T}}\mathbf{U}}\right], \boldsymbol{t} \in \mathbb{R}^n_+. \tag{3}$$

For ease of exposition, we will also refer to $\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})$ as the Laplace transform of $\mathbf{U}$.

II. Poisson Noise Model

Let $\mathbf{Y} \in \mathbb{Z}^k_+$ and $\mathbf{X} \in \mathbb{R}^n$. We say that $\mathbf{Y}$ is an output of a system with Poisson noise, if $\mathbf{Y}|\mathbf{X} = \boldsymbol{x}$ follows a Poisson distribution, that is,

$$P_{\mathbf{Y}|\mathbf{X}}(\mathrm{y}|\boldsymbol{x}) = \prod_{i=1}^k P_{Y_i|\mathbf{X}}(y_i|\boldsymbol{x}) \tag{4}$$

where

$$P_{Y_i|\mathbf{X}}(y_i|\boldsymbol{x}) = \frac{1}{y_i!}([\mathsf{A}\boldsymbol{x}]_i + \lambda_i)^{y_i} \mathrm{e}^{-([\mathsf{A}\boldsymbol{x}]_i+\lambda_i)}, \tag{5}$$

$\mathsf{A} \in \mathbb{R}^{k\times n}$ and $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_k]^{\mathsf{T}} \in \mathbb{R}^k_+$. In (5) we use the convention that $0^0 = 1$.

Using the terminology of laser communications, we refer to A as the *intensity matrix* and $\boldsymbol{\lambda}$ as the *dark current* vector. Moreover, we assume that the matrix A must satisfy the following non-negativity preserving constraint:

$$\mathsf{A}\boldsymbol{x} \in \mathbb{R}^k_+, \; \forall \boldsymbol{x} \in \mathbb{R}^n_+. \tag{6}$$

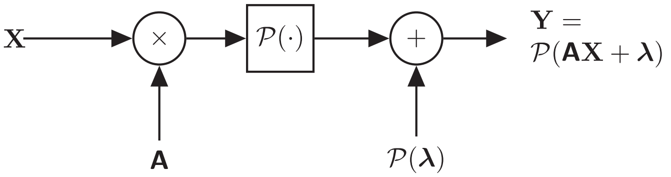


Fig. 1. The vector Poisson noise channel with the input $\mathbf{X}$, the intensity matrix A and the dark current parameter $\boldsymbol{\lambda}$.

The random transformation of the input random variable $\mathbf{X}$ to an output random variable $\mathbf{Y}$ by the channel in (4) is denoted by

$$\mathbf{Y} = \mathcal{P}(\mathsf{A}\mathbf{X} + \boldsymbol{\lambda}). \tag{7}$$

The transformation in (7) is depicted in Fig. 1.

III. Main Results

This section presents our main result pertaining to the linearity properties of the conditional expectation $\mathbb{E}[\mathbf{X}|\mathbf{Y} = \boldsymbol{y}]$. Specifically, our interest lies in answering various questions of optimality of linear estimators such as:

1) Under what prior distribution on $\mathbf{X}$ are linear estimators optimal for squared error loss and Bregman divergence[1] loss? Since the conditional expectation is an optimal estimator for the aforementioned loss functions, this is equivalent to asking when the conditional expectation is a linear function of $\boldsymbol{y}$.
2) Which linear estimators are realizable from $\mathbb{E}[\mathbf{X}|\mathbf{Y}]$? That is, given that $\mathbb{E}[\mathbf{X}|\mathbf{Y}] = \mathsf{H}\mathbf{Y} + \boldsymbol{c}$, what values of the matrix H and vector $\boldsymbol{c}$ are permitted?
3) If the linear estimators are approximately optimal, can we say something about the prior distribution of $\mathbf{X}$? In other words, we are looking for a quantitative refinement of 1).

Questions 1) and 2) are answered in Theorem 1 and Corollary 1, and question 3) is addressed in Theorem 2.

A. Necessary and Sufficient Conditions for Linearity

Our first result is the following theorem, the proof of which can be found in Section IV.

Theorem 1: Suppose that $\mathbf{Y} = \mathcal{P}(\mathbf{U})$ where $\mathbf{U}$ is a non-degenerate[2] random vector. Then,

$$\mathbb{E}[\mathbf{U}|\mathbf{Y} = \boldsymbol{y}] = \mathsf{H}\boldsymbol{y} + \boldsymbol{c}, \forall \boldsymbol{y} \in \mathbb{Z}^n_+ \tag{8}$$

if and only if

$$P_{\mathbf{U}} = \prod_{i=1}^n \mathsf{Gam}\left(\theta_i, \alpha_i\right). \tag{9}$$

In this case

[1] Let $\phi : \Omega \to \mathbb{R}$ be a *continuously-differentiable* and *a strictly convex* function defined on a *closed convex* set $\Omega \subseteq \mathbb{R}^n$. The Bregman divergence between u and v, associated with the function ϕ, is defined as $\ell_\phi(u, v) = \phi(u) - \phi(v) - \langle u - v, \nabla\phi(v)\rangle$.

[2] A random vector is said to be degenerate if its covariance of matrix is not full rank.

- H is diagonal with entries $h_{ii} = \dfrac{1}{1+\theta_i}$
- $c_i = \alpha_i h_{ii} = \dfrac{\alpha_i}{1+\theta_i}$

Note that $0 < h_{ii} < 1$ and $c_i > 0$ for all $i \in [1:n]$.

B. Quantitative Refinement of Theorem 1

In this section, a quantitative refinement of Theorem 1 is shown. Namely, it is shown that if the conditional mean estimator is close to a linear function in a mean squared error sense, then the prior distribution must be close to a product gamma distribution in terms of their Laplace transforms.

Theorem 2: Let H and $\boldsymbol{c}$ be as in Theorem 1 and let $\mathcal{L}_{\mathsf{G}}$ denote the Laplace transform of the product gamma distribution in (9). Assume that $\mathbf{Y} = \mathcal{P}(\mathbf{U})$ for some $\mathbf{U} \in \mathbb{R}_+^n$ and that

$$\mathbb{E}\left[\|\mathbb{E}[\mathbf{U}|\mathbf{Y}] - (\mathsf{H}\mathbf{Y} + \boldsymbol{c})\|^2\right] \le \epsilon \tag{10}$$

for some $\epsilon \ge 0$. Then,

$$\sup_{\boldsymbol{t}\in\mathbb{R}_+^n} \frac{|\mathcal{L}_{\mathbf{U}}(\boldsymbol{t}) - \mathcal{L}_{\mathsf{G}}(\boldsymbol{t})|}{\|\boldsymbol{t}\|} \le \frac{\sqrt{\epsilon}}{1 - \max_k h_{kk}}, \tag{11}$$

where $\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})$ is the Laplace transform of $\mathbf{U}$.

The proof of Theorem 2 is presented in Section V.

IV. Proof of Theorem 1

We first establish conditions on $\boldsymbol{c}$ and H under which the equality is possible.

A. Conditions on $\boldsymbol{c}$

To establish such conditions we need the following representation of the conditional expectation.

Lemma 1: Let $P_{\mathbf{Y}}$ denote the probability mass function of $\mathbf{Y}$. Then, for $\boldsymbol{y} \in \mathbb{Z}_+^n$

$$\mathbb{E}[\mathbf{U}|\mathbf{Y} = \boldsymbol{y}] = (\mathsf{diag}(\boldsymbol{y}) + \mathsf{I})\,\frac{\Delta P_{\mathbf{Y}}(\boldsymbol{y})}{P_{\mathbf{Y}}(\boldsymbol{y})}, \tag{12}$$

where

$$[\Delta P_{\mathbf{Y}}(\boldsymbol{y})]_i = P_{\mathbf{Y}}(\boldsymbol{y} + \mathbf{1}_i),\ i \in [1:n]. \tag{13}$$

The scalar version of Lemma 1 has been shown in [17] and in [18] and the vector version has been shown in [19, Lemma 3] and [9, Lemma 3].

We proceed to show that every element of $\boldsymbol{c}$ must be strictly positive. Choosing $\boldsymbol{y} = \mathbf{0}$ and combining (8) with (12) implies that

$$\boldsymbol{c} = \frac{\Delta P_{\mathbf{Y}}(\mathbf{0})}{P_{\mathbf{Y}}(\mathbf{0})}, \tag{14}$$

or equivalently for all i

$$c_i = \frac{P_{\mathbf{Y}}(\mathbf{0} + \mathbf{1}_i)}{P_{\mathbf{Y}}(\mathbf{0})} = \frac{\mathbb{E}\left[U_i \mathrm{e}^{-\sum_{i=1}^n U_i}\right]}{\mathbb{E}\left[\mathrm{e}^{-\sum_{i=1}^n U_i}\right]}. \tag{15}$$

The above is zero if and only if $U_i = 0$ and is positive otherwise.

B. Conditions on H

We now proceed to study properties of H. First, by combining (8) with (12), we have

$$\frac{\Delta P_{\mathbf{Y}}(\boldsymbol{y})}{P_{\mathbf{Y}}(\boldsymbol{y})} = (\mathsf{diag}(\boldsymbol{y}) + \mathsf{I})^{-1}(\mathsf{H}\boldsymbol{y} + \boldsymbol{c}) \tag{16}$$

$$= (\mathsf{diag}(\boldsymbol{y}) + \mathsf{I})^{-1}\mathsf{H}\boldsymbol{y} + (\mathsf{diag}(\boldsymbol{y}) + \mathsf{I})^{-1}\boldsymbol{c}, \tag{17}$$

which equivalently can be written as

$$\frac{P_{\mathbf{Y}}(\boldsymbol{y} + \mathbf{1}_i)}{P_{\mathbf{Y}}(\boldsymbol{y})} = \frac{1}{y_i + 1}\sum_{j=1} h_{ij} y_j + \frac{c_i}{y_i + 1}, \forall i \in [1:n]. \tag{18}$$

Observe that every entry of $\frac{\Delta P_{\mathbf{Y}}(\boldsymbol{y})}{P_{\mathbf{Y}}(\boldsymbol{y})}$ is non-negative. Therefore, for every i we have the following inequality:

$$0 \le \frac{1}{y_i + 1}\sum_{j=1} h_{ij} y_j + \frac{c_i}{y_i + 1}, \forall \boldsymbol{y} \in \mathbb{Z}_+^n, \tag{19}$$

where h_{ij} is the (i,j) element of H. Since $\boldsymbol{y}$ can be chosen arbitrary in (19), taking limits along all possible paths as y_i's go to infinity we arrive at

$$0 \le h_{ii} + \sum_{j\in S} h_{ij}, \forall i \text{ and } \forall S \subset [1:n] \setminus i. \tag{20}$$

In particular, by selecting S to be an empty set we arrive at the conclusion that $0 \le h_{ii}, \forall i$. To see that $h_{ii} \neq 0$, consider

$$\mathbb{E}[U_i|\mathbf{Y} = \mathbf{0} + y_i\mathbf{1}_i] = h_{ii}y_i + c_i, \forall y \in \mathbb{Z}_+. \tag{21}$$

Therefore, h_{ii} can only be zero if U_i is a constant.

Next, using (18) and summing over y_i we have that

$$\sum_{y_i=0}^{k} (y_i + 1) P_{\mathbf{Y}}(\boldsymbol{y} + \mathbf{1}_i) = \sum_{y_i=0}^{k}\left(\sum_{j=1} h_{ij} y_j + c_i\right) P_{\mathbf{Y}}(\boldsymbol{y}), \tag{22}$$

or, equivalently, by doing a change of variable on the left side of (22),

$$\mathbb{E}[Y_i 1_{\{Y_i \le k+1\}}|\mathbf{Y}_{-i} = \boldsymbol{y}_{-i}]$$
$$= \mathbb{E}\left[\left(\sum_{j=1} h_{ij} Y_j + c_i\right) 1_{\{Y_i \le k\}}|\mathbf{Y}_{-i} = \boldsymbol{y}_{-i}\right], \tag{23}$$

where $\mathbf{Y}_{-i}$ is $\mathbf{Y}$ with the i-th element removed. Now by choosing $\boldsymbol{y}_{-i} = \mathbf{0}$ and re-arranging the terms we have that

$$h_{ii} = \frac{\mathbb{E}[Y_i 1_{\{Y_i \le k+1\}}|\mathbf{Y}_{-i} = \mathbf{0}] - c_i\mathbb{E}\left[1_{\{Y_i \le k\}}|\mathbf{Y}_{-i} = \mathbf{0}\right]}{\mathbb{E}[Y_i 1_{\{Y_i \le k\}}|\mathbf{Y}_{-i} = \mathbf{0}]}, \tag{24}$$

for all k. Now taking k to infinity and using the fact that $c_i > 0$, it immediately follows that $h_{ii} < 1$.

The above discussion shows that $0 < h_{ii} < 1, \forall i$. We now proceed to show that H is invertible. To that end, we need the following lemma shown in Appendix A.

Lemma 2: For $\boldsymbol{y} \in \mathbb{Z}_+^n$

$$\begin{aligned} &[\mathsf{Var}(\mathbf{U}|\mathbf{Y}=\boldsymbol{y})]_{ij} \\ &= \mathbb{E}[U_i|\mathbf{Y}=\boldsymbol{y}]\left(\mathbb{E}[U_j|\mathbf{Y}=\boldsymbol{y}+\mathbf{1}_i] - \mathbb{E}[U_j|\mathbf{Y}=\boldsymbol{y}]\right). \quad (25) \end{aligned}$$

Now by using Lemma 2 and taking $\mathbb{E}[\mathbf{U}|\mathbf{Y}=\boldsymbol{y}] = \mathsf{H}\boldsymbol{y}+\boldsymbol{c}$ we have that

$$\mathsf{Var}(\mathbf{U}|\mathbf{Y}=\mathbf{0}) = \boldsymbol{c}\mathsf{1}^{\mathsf{T}} \odot \mathsf{H}^{\mathsf{T}} \quad (26)$$

where $\odot$ denotes the element-wise product (i.e., Hadamard product). Now using an elementary rank bound for the element-wise product, and the fact that for non-degenerate random vectors $\mathsf{Var}(\mathbf{U}|\mathbf{Y}=\mathbf{0})$ is a positive definite matrix, we have that

$$\begin{aligned} n &= \mathsf{Rank}\left(\mathsf{Var}(\mathbf{U}|\mathbf{Y}=\mathbf{0})\right) && (27)\\ &= \mathsf{Rank}\left(\boldsymbol{c}\mathsf{1}^{\mathsf{T}} \odot \mathsf{H}^{\mathsf{T}}\right) && (28)\\ &\le \mathsf{Rank}\left(\boldsymbol{c}\mathsf{1}^{\mathsf{T}}\right)\mathsf{Rank}\left(\mathsf{H}^{\mathsf{T}}\right) && (29)\\ &= \mathsf{Rank}\left(\mathsf{H}^{\mathsf{T}}\right) && (30)\\ &\le n. && (31) \end{aligned}$$

Therefore, H has full rank and is invertible.

We now proceed to show that H must be a diagonal matrix. The following lemma will be extensively used in this proof and the proof of Theorem 2.

Lemma 3: Let $\mathbf{Y} = \mathcal{P}(\mathbf{U})$ and suppose that (8) holds. Then,

$$\mathbb{E}\left[\left(\mathbf{U} - (\mathsf{H}\mathbf{Y}+\boldsymbol{c})\right)\mathrm{e}^{-\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}}\right] = \mathbf{0} \quad (32)$$

for all $\boldsymbol{t} \in \mathbb{R}_+^n$. Moreover, for all $\boldsymbol{t} \in \mathbb{R}_+^n$

$$\begin{aligned} &\mathbb{E}\left[\left(\mathbf{U} - (\mathsf{H}\mathbf{Y}+\boldsymbol{c})\right)\mathrm{e}^{-\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}}\right] \\ &= -(\mathsf{H}(\mathsf{diag}(\mathbf{s})-\mathsf{I})+\mathsf{I})\nabla_{\mathbf{s}}\mathcal{L}_{\mathbf{U}}(\mathbf{s}) - \boldsymbol{c}\mathcal{L}_{\mathbf{U}}(\mathbf{s}), \quad (33) \end{aligned}$$

where $s_m = 1-\mathrm{e}^{-t_m}, m = 1,\dots,n$.

Proof: The proof of (32) follows from the orthogonality principle. To show (33) we need to compute the following terms:

$$\mathbb{E}\left[\mathbf{U}\mathrm{e}^{-\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}}\right], \mathbb{E}\left[\mathrm{e}^{-\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}}\right] \text{ and } \mathbb{E}\left[\mathbf{Y}\mathrm{e}^{-\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}}\right]. \quad (34)$$

Also, recall that the Laplace transform of a distribution of a scalar Poisson random variable W with the parameter λ is given by

$$\mathcal{L}_W(t) = \mathrm{e}^{\lambda v(t)}, \quad (35)$$

where $v(t) = (\mathrm{e}^{-t}-1)$.

Now, first,

$$\begin{aligned} \mathbb{E}\left[\mathbf{U}\mathrm{e}^{\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}}\right] &= \mathbb{E}\left[\mathbf{U}\mathbb{E}\left[\mathrm{e}^{\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}} \mid \mathbf{U}\right]\right] && (36)\\ &= \mathbb{E}\left[\mathbf{U}\prod_{m=1}^{n}\mathbb{E}\left[\mathrm{e}^{t_m Y_m} \mid U_m\right]\right] && (37)\\ &= \mathbb{E}\left[\mathbf{U}\prod_{m=1}^{n}\mathrm{e}^{v(t_m)U_m}\right] && (38)\\ &= \mathbb{E}\left[\mathbf{U}\mathrm{e}^{-\mathbf{s}^{\mathsf{T}}\mathbf{U}}\right] && (39)\\ &= \nabla_{\mathbf{s}}\mathcal{L}_{\mathbf{U}}(\mathbf{s}), && (40) \end{aligned}$$

where (38) follows by using the Laplace transform of a scalar Poisson distribution. Second, using similar steps, we have

$$\begin{aligned} \mathbb{E}\left[\mathrm{e}^{-\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}}\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathrm{e}^{-\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}} \mid \mathbf{U}\right]\right] && (41)\\ &= \mathbb{E}\left[\prod_{i=m}^{n}\mathrm{e}^{v(t_m)U_i}\right] && (42)\\ &= \mathbb{E}\left[\mathrm{e}^{-\mathbf{s}^{\mathsf{T}}\mathbf{U}}\right] && (43)\\ &= \mathcal{L}_{\mathbf{U}}(\mathbf{s}). && (44) \end{aligned}$$

Third,

$$\begin{aligned} \mathbb{E}\left[\mathbf{Y}\mathrm{e}^{-\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}}\right] &= -\nabla_{\boldsymbol{t}}\mathbb{E}\left[\mathrm{e}^{-\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}}\right] && (45)\\ &= -\nabla_{\boldsymbol{t}}\mathbb{E}\left[\prod_{m=1}^{n}\mathrm{e}^{v(t_m)U_m}\right] && (46)\\ &= -\nabla_{\boldsymbol{t}}\mathbb{E}\left[\mathrm{e}^{-\mathbf{s}^{\mathsf{T}}\mathbf{U}}\right] && (47)\\ &= \mathbb{E}\left[\nabla_{\boldsymbol{t}}\mathbf{s}^{\mathsf{T}}\mathbf{U}\mathrm{e}^{-\mathbf{s}^{\mathsf{T}}\mathbf{U}}\right] && (48)\\ &= \mathbb{E}\left[(\mathsf{I}-\mathsf{diag}(\mathbf{s}))\mathbf{U}\mathrm{e}^{-\mathbf{s}^{\mathsf{T}}\mathbf{U}}\right] && (49)\\ &= (\mathsf{diag}(\mathrm{s})-\mathsf{I})\nabla_{\mathbf{s}}\mathcal{L}_{\mathbf{U}}(\mathbf{s}), && (50) \end{aligned}$$

where we have used that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t_m}s_m U_m &= \frac{\mathrm{d}}{\mathrm{d}t_m}(1-\mathrm{e}^{-t_m})U_m && (51)\\ &= \mathrm{e}^{-t_m}U_m && (52)\\ &= (1-s)U_m. && (53) \end{aligned}$$

Combining (40), (44) and (50) we arrive at

$$\begin{aligned} &\mathbb{E}\left[\left(\mathbf{U} - (\mathsf{H}\mathbf{Y}+\boldsymbol{c})\right)\mathrm{e}^{-\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}}\right] && (54)\\ &= \nabla_{\mathbf{s}}\mathcal{L}_{\mathbf{U}}(\mathrm{s}) - \mathsf{H}(\mathsf{diag}(\mathrm{s})-\mathsf{I})\nabla_{\mathbf{s}}\mathcal{L}_{\mathbf{U}}(\mathrm{s}) - \boldsymbol{c}\mathcal{L}_{\mathbf{U}}(\mathrm{s}) && (55)\\ &= -(\mathsf{H}\mathsf{diag}(\mathrm{s}) + (\mathsf{I}-\mathsf{H}))\nabla_{\mathbf{s}}\mathcal{L}_{\mathbf{U}}(\mathrm{s}) - \boldsymbol{c}\mathcal{L}_{\mathbf{U}}(\mathrm{s}). && (56) \end{aligned}$$

This concludes the proof. ■

To present the solution to the differential equation in (33) we need the following lemma.

First, by using that H is invertible it follows that

$$\frac{\nabla_{\mathbf{s}}\mathcal{L}_{\mathbf{U}}(\mathbf{s})}{\mathcal{L}_{\mathbf{U}}(\mathbf{s})} = -\left(\mathsf{H}^{-1}(\mathsf{I}-\mathsf{H}) + \mathsf{diag}(\mathbf{s})\right)^{-1}\mathsf{H}^{-1}\boldsymbol{c}, \quad (57)$$

which can further be simplified to

$$\nabla g(\mathbf{s}) = \left(\mathsf{H}^{-1}(\mathsf{I}-\mathsf{H}) + \mathsf{diag}(\mathbf{s})\right)^{-1}\mathsf{H}^{-1}\boldsymbol{c}, \quad (58)$$

where $g(\mathbf{s}) = \log(\mathcal{L}_{\mathbf{U}}(\mathbf{s}))$.

Next it is shown that (58) has a solution only if H is a diagonal matrix and the solution is characterized.

Lemma 4: For $\mathbf{0} \prec \mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{b} \in \mathbb{R}^n$ where $\mathbf{b}$ is assumed to have all positive entries. The system

$$\nabla g(\mathbf{s}) = -(\mathbf{A} + \mathsf{diag}(\mathbf{s}))^{-1}\mathbf{b},\ g(\mathbf{0}) = \mathbf{0}, \tag{59}$$

has a solution only if $\mathbf{A}$ is a diagonal matrix with a solution given by

$$g(\mathbf{s}) = \sum_{i=1}^{n} b_i \log\left(1 + \frac{s_i}{A_{ii}}\right). \tag{60}$$

Proof: We first find the Hessian matrix of $f(\mathbf{s}) = \nabla g(\mathbf{s})$. Let

$$\mathsf{C} = \mathbf{A} + \mathsf{diag}(\mathbf{s}), \tag{61}$$

$$\mathsf{S} = \mathsf{diag}(\mathbf{s}), \tag{62}$$

$$\mathsf{S}f = \mathsf{diag}(f)\mathbf{s}. \tag{63}$$

Then, the differential is given by

$$\partial f = \partial \mathsf{C}^{-1}\mathbf{b} \tag{64}$$

$$= -\mathsf{C}^{-1}(\partial \mathsf{C})\mathsf{C}^{-1}\mathbf{b} \tag{65}$$

$$= -\mathsf{C}^{-1}(\partial \mathsf{C})f \tag{66}$$

$$= -\mathsf{C}^{-1}(\partial \mathsf{S})f \tag{67}$$

$$= -\mathsf{C}^{-1}\mathsf{diag}(f)\partial \mathbf{s}. \tag{68}$$

Hence,

$$\frac{\partial f}{\partial \mathbf{s}} = -\mathsf{C}^{-1}\mathsf{diag}(f) = -\mathsf{C}^{-1}\mathsf{diag}((\mathbf{A} + \mathsf{diag}(\mathbf{s}))^{-1}\mathbf{b}). \tag{69}$$

Therefore, the Hessian matrix of g is given by

$$\nabla^2 g(\mathbf{s}) = -(\mathbf{A} + \mathsf{diag}(\mathbf{s}))^{-1}\mathsf{diag}\left((\mathbf{A} + \mathsf{diag}(\mathbf{s}))^{-1}\mathbf{b}\right). \tag{70}$$

Note that the Hessian matrix must be symmetric. Next, it is shown that in order for the Hessian to be symmetric $\mathbf{A}$ must be a diagonal matrix.

Let $\tilde{\mathbf{A}} = (\mathbf{A} + \mathsf{diag}(\mathbf{s}))^{-1}$ and choose $\mathbf{s}$ such that

$$\tilde{\mathbf{b}} = \tilde{\mathbf{A}}\mathbf{b} = (\mathbf{A} + \mathsf{diag}(\mathbf{s}))^{-1}\mathbf{b} \tag{71}$$

has distinct elements all of which are non-zero. Note that this is possible in view of the assumption that $\mathbf{b}$ has non-zero entries.

Next, observe that if $\tilde{\mathbf{A}}\mathsf{diag}(\tilde{\mathbf{b}})$ is symmetric, then $\tilde{\mathbf{A}}$ must be symmetric. This follows by letting $\tilde{\mathsf{C}} = \tilde{\mathbf{A}}\mathsf{diag}(\tilde{\mathbf{b}})$ and observing that $\tilde{\mathbf{A}} = \tilde{\mathsf{C}}\mathsf{diag}(\tilde{\mathbf{b}})^{-1}$ is symmetric. The symmetry of $\tilde{\mathbf{A}}$ implies that

$$\tilde{\mathbf{A}}\mathsf{diag}(\tilde{\mathbf{b}}) = \mathsf{diag}(\tilde{\mathbf{b}})\tilde{\mathbf{A}}^{\mathsf{T}} = \mathsf{diag}(\tilde{\mathbf{b}})\tilde{\mathbf{A}}. \tag{72}$$

In other words, $\tilde{\mathbf{A}}$ and $\mathsf{diag}(\tilde{\mathbf{b}})$ commute. However, if all elements of a diagonal matrix are distinct, then it commutes only with a diagonal matrix. Therefore, $\tilde{\mathbf{A}}$ is a diagonal matrix. This implies that for the Hessian to be symmetric $\mathbf{A}$ must be a diagonal matrix.

Since $\mathbf{A}$ is diagonal, the solution is obtained by an application of the fundamental theorem of calculus for line integrals: for a function f and a smooth curve $\boldsymbol{r}(t)$ we have

$$\int_a^b \nabla f(\boldsymbol{r}(t)) \cdot \dot{\boldsymbol{r}}(t)\mathrm{d}t = f(\boldsymbol{r}(b)) - f(\boldsymbol{r}(a)). \tag{73}$$

Applying (73) to (59) with a choice of $\boldsymbol{r}(t) = (1-t)\mathbf{0} + t\mathbf{s}$, $t \in (0,1)$, we have that

$$g(\mathbf{s}) = -\int_0^1 (\mathbf{A} + \mathsf{diag}(\mathbf{s})t)^{-1}\mathbf{b}\cdot \mathbf{s}\mathrm{d}t \tag{74}$$

$$= -\mathbf{s}^{\mathsf{T}}\int_0^1 (\mathbf{A} + \mathsf{diag}(\mathbf{s})t)^{-1}\mathrm{d}t\,\mathbf{b} \tag{75}$$

$$= -\mathbf{s}^{\mathsf{T}}\mathsf{diag}\left(\left[\frac{\log(1 + \frac{s_k}{A_{kk}})}{s_k}\right]_k\right)\mathbf{b} \tag{76}$$

$$= -\sum_{k=1}^{n} b_k \log\left(1 + \frac{s_k}{A_{kk}}\right). \tag{77}$$

■

Setting $\mathbf{A} = \mathbf{H}^{-1}(\mathbf{I} - \mathbf{H})$ and $\mathbf{b} = \mathbf{H}^{-1}\boldsymbol{c}$ in Lemma 4 and using that $g(\mathbf{s}) = \log(\mathcal{L}_{\mathbf{U}}(\mathbf{s}))$ we arrive at the following form for the Laplace transform of the distribution of $\mathbf{U}$:

$$\mathcal{L}_{\mathbf{U}}(\mathbf{s}) = \prod_{k=1}^{n} \frac{1}{\left(1 + \frac{h_{kk}s_k}{1-h_{kk}}\right)^{\frac{h_{kk}}{c_k}}}, \tag{78}$$

which is the Laplace transform of a product of Gamma distributions. This concludes the proof.

V. Proof Theorem 2

Let the characteristic function of the product gamma distribution be denoted by

$$\mathcal{L}_{\mathsf{G}}(\boldsymbol{t}) = \prod_{k=1}^{n}\left(1 + \frac{t_k}{\alpha_k}\right)^{-\theta_k}. \tag{79}$$

The following result, which is a generalization of the scalar result in [12], will be useful.

Lemma 5: Let $\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})$ be the Laplace transform of a non-negative random vector $\mathbf{U}$ and let

$$\mathbf{A} = \mathsf{diag}^{-1}\left([\alpha_1, \ldots, \alpha_n]^{\mathsf{T}}\right), \tag{80}$$

$$\tilde{\boldsymbol{c}} = \left[\frac{\theta_1}{\alpha_1}, \ldots, \frac{\theta_k}{\alpha_k}\right]^{\mathsf{T}}. \tag{81}$$

Then, for every $\boldsymbol{t} \in \mathbb{R}_+^n$

$$|\mathcal{L}_{\mathbf{U}}(\boldsymbol{t}) - \mathcal{L}_{\mathsf{G}}(\boldsymbol{t})| \\ \leq \|\boldsymbol{t}\| \sup_{\boldsymbol{t}\in\mathbb{R}_+^n} \|(\mathbf{I} + \mathbf{A}\mathsf{diag}(\boldsymbol{t}))\nabla\phi_{\mathbf{U}}(\boldsymbol{t}) + \tilde{\boldsymbol{c}}\phi_{\mathbf{U}}(\boldsymbol{t})\|. \tag{82}$$

Proof: First, note that

$$\frac{\partial}{\partial t_k}\frac{1}{\mathcal{L}_{\mathsf{G}}(\boldsymbol{t})} = \frac{\theta_k}{\alpha_k}\frac{1}{\left(1 + \frac{t_k}{\alpha_k}\right)\mathcal{L}_{\mathsf{G}}(\boldsymbol{t})}, \tag{83}$$

and hence

$$\frac{\partial}{\partial t_k}\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})\frac{1}{\mathcal{L}_{\mathsf{G}}(\boldsymbol{t})} \\ = \frac{\partial}{\partial t_k}\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})\frac{1}{\mathcal{L}_{\mathsf{G}}(\boldsymbol{t})} + \mathcal{L}_{\mathbf{U}}(\boldsymbol{t})\frac{\partial}{\partial t_k}\frac{1}{\mathcal{L}_{\mathsf{G}}(\boldsymbol{t})} \tag{84}$$

$$= \frac{1}{\left(1+\frac{t_k}{\alpha_k}\right)\mathcal{L}_{\mathsf{G}}(\boldsymbol{t})} \left(\left(1+\frac{t_k}{\alpha_k}\right)\frac{\partial}{\partial t_k}\mathcal{L}_{\mathbf{U}}(\boldsymbol{t}) + \frac{\theta_k}{\alpha_k}\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})\right). \tag{85}$$

Therefore, the gradient can be upper bounded as

$$\left\|\nabla\left(\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})\frac{1}{\mathcal{L}_{\mathsf{G}}(\boldsymbol{t})}\right)\right\|$$

$$= \frac{\left\|(\mathsf{I}+\mathsf{A}\mathrm{diag}(\boldsymbol{t}))^{-1}\left((\mathsf{I}+\mathsf{A}\mathrm{diag}(\boldsymbol{t}))\nabla\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})+\tilde{\boldsymbol{c}}\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})\right)\right\|}{\mathcal{L}_{\mathsf{G}}(\boldsymbol{t})} \tag{86}$$

$$\leq \frac{\left\|(\mathsf{I}+\mathsf{A}\mathrm{diag}(\boldsymbol{t}))^{-1}\right\|_{*}\|(\mathsf{I}+\mathsf{A}\mathrm{diag}(\boldsymbol{t}))\nabla\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})+\tilde{\boldsymbol{c}}\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})\|}{\mathcal{L}_{\mathsf{G}}(\boldsymbol{t})} \tag{87}$$

where $\|\cdot\|_{\star}$ denotes the operator norm.

Next, recall that the operator norm of a diagonal matrix is given by the maximal element and

$$\left\|(\mathsf{I}+\mathsf{A}\mathrm{diag}(\boldsymbol{t}))^{-1}\right\|_{*} = \max_{k\in[1:n]}\left|1+\frac{t_k}{\alpha_k}\right|^{-1}$$

$$= \left(1+\min_{k\in[1:n]}\frac{t_k}{\alpha_k}\right)^{-1}. \tag{88}$$

Now let $r(\tau)=\tau\boldsymbol{t}$ and observe the following sequence of steps:

$$|\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})-\mathcal{L}_{\mathsf{G}}(\boldsymbol{t})|$$

$$= \mathcal{L}_{\mathsf{G}}(\boldsymbol{t})\left|\frac{\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})}{\mathcal{L}_{\mathsf{G}}(\boldsymbol{t})}-1\right| \tag{89}$$

$$= \mathcal{L}_{\mathsf{G}}(\boldsymbol{t})\left|\int_0^1 \nabla\frac{\mathcal{L}_{\mathbf{U}}(\boldsymbol{r}(\tau))}{\mathcal{L}_{\mathsf{G}}(\boldsymbol{r}(\tau))}\cdot\dot{\boldsymbol{r}}(\tau)\mathrm{d}\tau\right| \tag{90}$$

$$\leq \mathcal{L}_{\mathsf{G}}(\boldsymbol{t})\int_0^1\left\|\nabla\frac{\mathcal{L}_{\mathbf{U}}(\boldsymbol{r}(\tau))}{\mathcal{L}_{\mathsf{G}}(\boldsymbol{r}(\tau))}\right\|\|\dot{\boldsymbol{r}}(\tau)\|\,\mathrm{d}\tau \tag{91}$$

$$\leq \|\boldsymbol{t}\|\sup_{\boldsymbol{t}\in\mathbb{R}_+^n}\|(\mathsf{I}+\mathsf{A}\mathrm{diag}(\boldsymbol{t}))\nabla\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})+\tilde{\boldsymbol{c}}\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})\|, \tag{92}$$

where (90) follows from the fundamental theorem of calculus for line integrals; (91) follows by using the Cauchy-Schwarz inequality; and (92) follows by using the bound in (87), the fact that $\mathcal{L}_{\mathsf{G}}(\tau\boldsymbol{t})$ is a decreasing function of τ, and

$$\int_0^1\frac{\|(\mathsf{I}+\mathsf{A}\mathrm{diag}(\tau\boldsymbol{t}))\nabla\mathcal{L}_{\mathbf{U}}(\tau\boldsymbol{t})+\tilde{\boldsymbol{c}}\mathcal{L}_{\mathbf{U}}(\tau\boldsymbol{t})\|}{\mathcal{L}_{\mathsf{G}}(\tau\boldsymbol{t})\left(1+\min_{k\in[1:n]}\frac{\tau t_k}{\alpha_k}\right)}\mathrm{d}\tau$$

$$\leq \frac{1}{\mathcal{L}_{\mathsf{G}}(\boldsymbol{t})}\int_0^1\frac{\|(\mathsf{I}+\mathsf{A}\mathrm{diag}(\tau\boldsymbol{t}))\nabla\mathcal{L}_{\mathbf{U}}(\tau\boldsymbol{t})+\tilde{\boldsymbol{c}}\mathcal{L}_{\mathbf{U}}(\tau\boldsymbol{t})\|}{\left(1+\min_{k\in[1:n]}\frac{\tau t_k}{\alpha_k}\right)}\mathrm{d}\tau$$

$$\leq \frac{\sup_{\boldsymbol{t}\in\mathbb{R}_+^n}\|(\mathsf{I}+\mathsf{A}\mathrm{diag}(\boldsymbol{t}))\nabla\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})+\tilde{\boldsymbol{c}}\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})\|}{\mathcal{L}_{\mathsf{G}}(\boldsymbol{t})}. \tag{93}$$

This concludes the proof. ■

With Lemma 5 at our disposal we are now ready to proof the main result. First, note that by Lemma 3 we have that

$$\|((\mathsf{I}-\mathsf{H})+\mathsf{H}\mathrm{diag}(\mathbf{s}))\nabla_{\mathbf{s}}\mathcal{L}_{\mathbf{U}}(\mathbf{s})+\boldsymbol{c}\mathcal{L}_{\mathbf{U}}(\mathbf{s})\|$$

$$= \left\|\mathbb{E}\left[(\mathbf{U}-(\mathsf{H}\mathbf{Y}+\boldsymbol{c}))\,\mathrm{e}^{-\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}}\right]\right\| \tag{94}$$

$$= \left\|\mathbb{E}\left[(\mathbb{E}[\mathbf{U}|\mathbf{Y}]-(\mathsf{H}\mathbf{Y}+\boldsymbol{c}))\,\mathrm{e}^{\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}}\right]\right.$$

$$\left.+\mathbb{E}\left[(\mathbf{U}-(\mathbb{E}[\mathbf{U}|\mathbf{Y}])\,\mathrm{e}^{-\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}}\right]\right\| \tag{95}$$

$$= \left\|\mathbb{E}\left[(\mathbb{E}[\mathbf{U}|\mathbf{Y}]-(\mathsf{H}\mathbf{Y}+\boldsymbol{c}))\,\mathrm{e}^{-\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}}\right]\right\| \tag{96}$$

$$\leq \mathbb{E}\left[\|\mathbb{E}[\mathbf{U}|\mathbf{Y}]-(\mathsf{H}\mathbf{Y}+\boldsymbol{c})\|\right] \tag{97}$$

$$\leq \sqrt{\mathbb{E}\left[\|\mathbb{E}[\mathbf{U}|\mathbf{Y}]-(\mathsf{H}\mathbf{Y}+\boldsymbol{c})\|^2\right]}, \tag{98}$$

where (95) follows by the orthogonality principle; (97) follows by using the modulus inequality and bound $\mathrm{e}^{-\boldsymbol{t}^{\mathsf{T}}\mathbf{Y}}\leq 1, \boldsymbol{t}\in\mathbb{R}_+^n$; and (98) follows by using Jensen's inequality.

Now by setting $\tilde{\boldsymbol{c}}=(\mathsf{I}-\mathsf{H})^{-1}\boldsymbol{c}$ and $\mathsf{A}=(\mathsf{I}-\mathsf{H})^{-1}\mathsf{H}$ in Lemma 5 we have that

$$|\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})-\mathcal{L}_{\mathsf{G}}(\boldsymbol{t})|$$

$$\leq \|\boldsymbol{t}\|\sup_{\boldsymbol{t}\in\mathbb{R}_+^n}\|(\mathsf{I}+\mathsf{A}\mathrm{diag}(\boldsymbol{t}))\nabla\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})+\tilde{\boldsymbol{c}}\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})\| \tag{99}$$

$$\leq \|\boldsymbol{t}\|\|(\mathsf{I}-\mathsf{H})^{-1}\|_{\star}$$

$$\cdot\sup_{\boldsymbol{t}\in\mathbb{R}_+^n}\|((\mathsf{I}-\mathsf{H})+\mathsf{H}\mathrm{diag}(\mathbf{t}))\nabla\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})+\boldsymbol{c}\mathcal{L}_{\mathbf{U}}(\boldsymbol{t})\| \tag{100}$$

$$\leq \|\boldsymbol{t}\|\|(\mathsf{I}-\mathsf{H})^{-1}\|_{\star}\sqrt{\mathbb{E}\left[\|\mathbb{E}[\mathbf{U}|\mathbf{Y}]-(\mathsf{H}\mathbf{Y}+\boldsymbol{c})\|^2\right]} \tag{101}$$

$$= \|\boldsymbol{t}\|\frac{\sqrt{\epsilon}}{1-\max_k h_{kk}}, \tag{102}$$

where (101) follows by using the bound in (98); and (102) using the fact that the operator norm of a diagonal matrix is given by the maximal element.

This concludes the proof.

VI. Conclusion

This section discusses implications of our results for the practically relevant model $\mathbf{Y}=\mathcal{P}(\mathsf{A}\mathbf{X}+\boldsymbol{\lambda})$, which explicitly takes into account the intensity matrix A and the dark current parameter $\boldsymbol{\lambda}$. In addition, we also compare the Poisson results obtained in this work to their Gaussian counterparts.

We begin by adopting Theorem 1 to the parametrization $\mathbf{Y}=\mathcal{P}(\mathsf{A}\mathbf{X}+\boldsymbol{\lambda})$. This is done by setting $\mathbf{U}=\mathsf{A}\mathbf{X}+\boldsymbol{\lambda}$ in Theorem 1.

Corollary 1: Suppose that $\mathbf{Y}=\mathcal{P}(\mathsf{A}\mathbf{X}+\boldsymbol{\lambda})$. Then,

$$\mathbb{E}[\mathbf{X}|\mathbf{Y}=\boldsymbol{y}]=\mathsf{C}\boldsymbol{y}+\mathbf{b}, \forall\boldsymbol{y}\in\mathbb{Z}_+^n \tag{103}$$

if and only if all of the following conditions hold:

- $\boldsymbol{\lambda}=\mathbf{0}$;
- AC is a diagonal matrix with $0<[\mathsf{AC}]_{ii}<1, \forall i\in[1:n]$;
- $\mathsf{A}\mathbf{b}$ is a vector of positive elements; and

- $P_{\mathbf{AX}} = \prod_{i=1}^{n} \mathsf{Gam}(\frac{1-[\mathbf{AC}]_{ii}}{[\mathbf{AC}]_{ii}}, \frac{[\mathbf{Ab}]_i}{[\mathbf{AC}]_{ii}})$

Proof: Let $\mathbf{U} = \mathbf{AX} + \boldsymbol{\lambda}$. By multiplying (103) by $\mathbf{A}$ and adding $\boldsymbol{\lambda}$ we have that

$$\mathbb{E}[\mathbf{U}|\mathbf{Y} = \boldsymbol{y}] = \mathbf{A}\mathbb{E}[\mathbf{X}|\mathbf{Y} = \boldsymbol{y}] + \boldsymbol{\lambda} = \mathbf{AC}\boldsymbol{y} + \mathbf{Ab} + \boldsymbol{\lambda}. \tag{104}$$

Next, note that the linearity of the conditional expectation implies that $\mathbf{U} = \mathbf{AX} + \boldsymbol{\lambda}$ is according with a product gamma distribution which has non-negative support. However, if $\boldsymbol{\lambda}$ has positive components, this would imply that $\mathbf{AX} = \mathbf{U} - \boldsymbol{\lambda}$ has negative components, which is not allowed under the Poisson model. Therefore, $\boldsymbol{\lambda}$ must be zero.

The rest of the argument follows from Theorem 1 by mapping $\mathbf{AC}$ to $\mathbf{H}$ and $\mathbf{Ab}$ to $\boldsymbol{c}$. ∎

A few comments are now in order.

A. The Case of a Non-Zero Dark Current

Somewhat regrettably Corollary 1 shows that the conditional expectation can only be linear if the dark current parameter is zero. To demonstrate the effect of the dark current we investigate a scalar case with an exponential distribution as a prior (i.e., a gamma distribution with $\theta = 1$).

Lemma 6: Let $Y = \mathcal{P}(aX + \lambda)$ and take X to be an exponential random variable of rate α. Then, for every $a > 0$ and $\lambda \geq 0$

$$\mathbb{E}[X|Y = k] = \frac{1}{a}\frac{(k+1)P_Y(k+1)}{P_Y(k)} - \frac{\lambda}{a}, \tag{105}$$

where

$$P_Y(0) = \frac{\alpha e^{-\lambda}}{\alpha + a}, \tag{106}$$

$$P_Y(k) = \frac{\Gamma(k+1,\lambda)}{\Gamma(k+1)} - \frac{\Gamma(k,\lambda)}{\Gamma(k)} + \frac{e^{\frac{\alpha}{a}\lambda}}{\left(1+\frac{\alpha}{a}\right)^k}\left(\frac{\Gamma\left(k, \lambda\left(\frac{\alpha}{a}+1\right)\right)}{\Gamma(k)} - \frac{\Gamma\left(k+1, \lambda\left(\frac{\alpha}{a}+1\right)\right)}{\Gamma(k+1)\left(1+\frac{\alpha}{a}\right)}\right), \tag{107}$$

where $\Gamma(\cdot,\cdot)$ is the upper incomplete gamma function.

Proof: (105) is a scalar version of Lemma 1. The proof of (106) and (107) follows by invoking standard integration techniques for exponential functions. ∎

The effect of the dark current parameter on the conditional expectation in the scalar case for an exponential random variable is shown in Fig. 2. Observe that the larger the dark current, the smaller the conditional expectation is. The interpretation here is that large values of dark current inflate the observed count at Y, and the estimator compensates by producing smaller estimates of X.

It is also interesting to compare the optimal linear estimator under the squared error loss to the conditional expectation. The former is given by

$$\widehat{X}(y) = cy + b, \tag{108}$$

$$c = \frac{a\mathbb{V}(X)}{a^2\mathbb{V}(X) + a\mathbb{E}[X] + \lambda}, \tag{109}$$

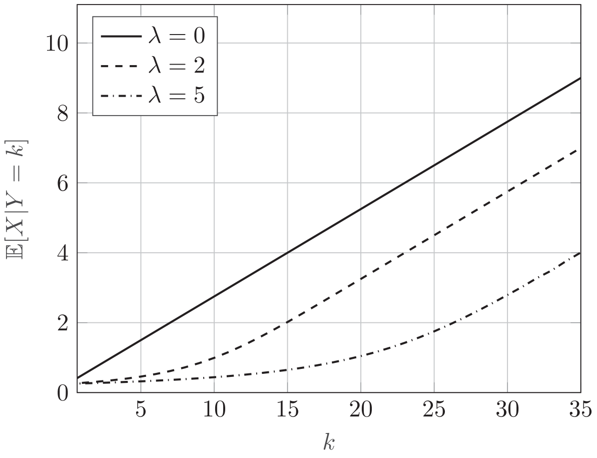


Fig. 2. Examples of conditional expectations for X distributed according to an exponential distribution with rate parameter $\alpha = 3$ and $a = 1$.

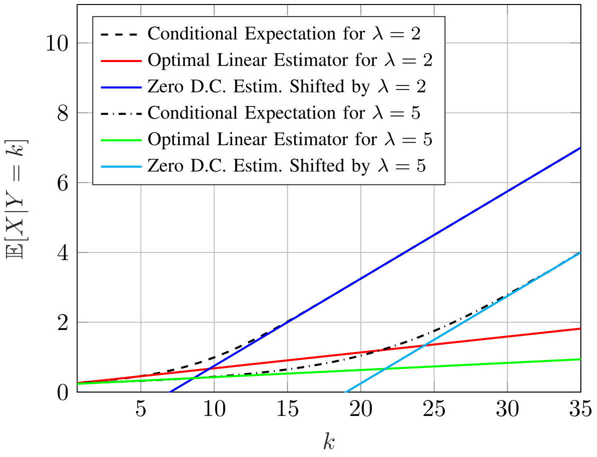


Fig. 3. Examples of conditional expectations and linear estimators for X distributed according to an exponential distribution with rate parameter $\alpha = 3$ and $a = 1$.

$$b = \mathbb{E}[X] - c(a\mathbb{E}[X] + \lambda). \tag{110}$$

Fig. 3 compares the conditional expectation to the optimal linear estimator for an exponential random variable and shows that the conditional expectation can be approximated by a piece-wise linear function. More specifically, Fig. 3 shows that the optimal linear estimator is a good approximation of the conditional expectation for small values of count, and the optimal zero dark current linear estimator shifted by the value of the dark current is a good approximation for large values of count.

B. On the Size of $\mathbf{A}$

Observe that according to Corollary 1 $\mathbf{AX}$ must have a product gamma distribution. The following scenarios can be encountered:

- $\mathbf{A}$ is full rank. In this case, the pdf of $\mathbf{X}$ is given by

$$f_{\mathbf{X}}(\boldsymbol{x}) = |\det(\mathbf{A})| f_{\mathbf{U}}(\mathbf{A}\boldsymbol{x}) \tag{111}$$

where $f_{\mathbf{U}}(\cdot)$ is the pdf of the product gamma distribution in Corollary 1.

- A is a 'fat' matrix (i.e., $k < n$). In this case, there are several distributions on $\mathbf{X}$ that result in a product gamma distribution; and
- A is a 'thin' matrix (i.e., $k > n$). In this case, in general, it is not possible to generate a product distribution.

C. Comparison to the Gaussian Noise Case

It is of some value to compare the result in the Poisson case to the Gaussian noise case. The objective is not to say which model is more useful, which clearly depends on the application, but rather to emphasize the difference in the behavior of the two models.

The Gaussian counterpart of Theorem 1, which is a well-known result (see for example [20, Lemma 5]), is given next.

Theorem 3: Suppose that $\mathsf{A} \in \mathbb{R}^{k\times n}$. Let $\mathbf{Y} = \mathsf{A}\mathbf{X} + \mathbf{Z}$ where $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathsf{I})$ are independent. Then,

$$\mathbb{E}[\mathbf{X}|\mathbf{Y} = \mathbf{y}] = \mathsf{H}\boldsymbol{y} + \boldsymbol{c}, \forall \boldsymbol{y} \in \mathbb{R}^n \tag{112}$$

if and only if $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathsf{K})$ such that

$$\mathsf{H} = \mathsf{K}\mathsf{A}^{\mathsf{T}}\left(\mathsf{A}\mathsf{K}\mathsf{A}^{\mathsf{T}} + \mathsf{I}\right)^{-1}, \tag{113}$$

$$\boldsymbol{c} = \boldsymbol{\mu} - \mathsf{H}\mathsf{A}\boldsymbol{\mu}. \tag{114}$$

In particular, $\mathsf{A}\mathbf{X} \sim \mathcal{N}(\mathsf{A}\mu, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = (\mathsf{I} - \mathsf{A}\mathsf{H})^{-1}\mathsf{A}\mathsf{H}$.

In Theorem 3, while the distribution on $\mathsf{A}\mathbf{X}$ is unique, the distribution on $\mathbf{X}$ may not be unique and depends on the dimensionality of A.

There are several high-level differences between Theorem 1 and Theorem 3. The restrictions on H in the Gaussian noise case are less severe than those in the Poisson noise case. First, H in the Gaussian noise case does not have to be a diagonal matrix while in Poisson noise case H has to be a diagonal matrix. Second, in the Poisson noise case the prior distribution has to be of a product form, but in the Gaussian noise case the prior can have arbitrary covariance matrix K and does not have to be a product distribution.

We now compare Theorem 2 to its Gaussian counterpart. Unlike in the Poisson noise case, where the prior distributions are supported on $\mathbb{R}_+^n$, in the Gaussian noise case the prior distribution are supported on $\mathbb{R}^n$. Therefore, it is more appropriated to provide a quantitative refinement in terms of the characteristic functions instead of the Laplace transforms.

However, to the best of our knowledge, for the Gaussian noise case there exists only a scalar counterpart of Theorem 2, which was shown in [21, Lemma 4]. In order to make a proper comparison, the following result provides a vector Gaussian generalization.

Theorem 4: Let H and $\boldsymbol{c}$ be as in Theorem 3. Denote by $\phi_{\mathsf{A}\mathbf{X}}(\boldsymbol{t})$, $\phi_{\mathbf{Z}}(\boldsymbol{t})$ and $\phi_{\mathbf{Y}}(\boldsymbol{t})$ the characteristic functions of $\mathsf{A}\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{Y}$, respectively. Assume that

$$\mathbb{E}\left[\|\mathbb{E}[\mathbf{X}|\mathbf{Y}] - (\mathsf{H}\mathbf{Y} + \boldsymbol{c})\|^2\right] \leq \epsilon, \tag{115}$$

for some $\epsilon \geq 0$. Then, for all $\boldsymbol{t} \in \mathbb{R}^k$

$$\frac{\left|\phi_{\mathsf{A}\mathbf{X}}(\boldsymbol{t}) - \mathrm{e}^{-\frac{\boldsymbol{t}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{t}}{2}}\right|}{\|\boldsymbol{t}\|} \leq \frac{\sqrt{\epsilon}\|\mathsf{A}\|_\star}{\sigma_{\min}\left(\mathsf{I} - \mathsf{A}\mathsf{H}\right)\phi_{\mathbf{Z}}\left(\boldsymbol{t}\right)} \tag{116}$$

where $\boldsymbol{\Sigma} = (\mathsf{I} - \mathsf{A}\mathsf{H})^{-1}\mathsf{A}\mathsf{H}$, $\|\mathsf{A}\|_\star$ is the operator norm of A, and $\sigma_{\min}(\mathsf{I} - \mathsf{A}\mathsf{H})$ is the smallest singular value of $\mathsf{I} - \mathsf{A}\mathsf{H}$. Consequently,

$$\sup_{\boldsymbol{t}\in\mathbb{R}^k} \frac{\left|\phi_{\mathbf{Y}}(\boldsymbol{t}) - \mathrm{e}^{-\frac{\boldsymbol{t}^{\mathsf{T}}(\boldsymbol{\Sigma}+\mathsf{I})\boldsymbol{t}}{2}}\right|}{\|\boldsymbol{t}\|} \leq \frac{\sqrt{\epsilon}\|\mathsf{A}\|_\star}{\sigma_{\min}\left(\mathsf{I} - \mathsf{A}\mathsf{H}\right)}. \tag{117}$$

Proof: See Appendix B. ■

It is interesting to compare the Poisson result in (11) to the Gaussian results in (116) and (117). First, note that in the Poisson noise case in (11), the control over the Laplace transform of the prior distributions is uniform over all $\boldsymbol{t}$. However, in the Gaussian noise case in (116), such a bound is not uniform over all $\boldsymbol{t}$. Therefore, the control over the proximity of the priors in terms of the proximity of estimators is stronger in Poisson noise. Finally, in the Gaussian noise case, we do get a uniform bound, but only a bound for the characteristic functions of the output $\mathbf{Y}$ as shown in (117).

D. Possible Applications

We briefly mention a few possible applications of our results:

- (Linear Minimum Mean Squared Error): Consider the minimum mean squared error (MMSE):

  $$\mathrm{mmse}(\mathbf{X}|\mathbf{Y}) = \mathbb{E}[\|\mathbf{X} - \mathbb{E}[\mathbf{X}|\mathbf{Y}]\|^2]. \tag{118}$$

  In general, computation of the MMSE can be a difficult task. Therefore, in order to provide performance guarantees, the MMSE is often upper bounded by the linear MMSE (i.e., the best error achievable with a linear estimator), that is,

  $$\mathrm{mmse}(\mathbf{X}|\mathbf{Y}) \leq \min_{\mathsf{H},\mathbf{c}} \mathbb{E}[\|\mathbf{X} - (\mathsf{H}\mathbf{Y} + \mathbf{c})\|^2]. \tag{119}$$

  Note that the linear MMSE has a closed-form expression and is relatively easy to evaluate. The result in Theorem 1 provides conditions on the triplet $(\mathbf{X}, \mathsf{H}, \mathbf{c})$ such that the bound in (119) is tight, and thus provides justification for such an analysis.
- (Least Favorable Distributions): Recall that the *least favorable prior distribution* (LFPD) is defined as

  $$P_{\mathbf{X}} \in \arg\max_{P_{\mathbf{X}}\in\mathcal{P}} \mathrm{mmse}(\mathbf{X}|\mathbf{Y}), \tag{120}$$

  where $\mathcal{P}$ is some set of distributions. The set $\mathcal{P}$ is defined by the practical constrains of the problem. For example, $\mathcal{P} = \{P_{\mathbf{X}} : \|\mathbf{X}\| \leq r\}$ where $r > 0$ is known as the restricted parameter estimation setting. LFPDs are important because $\mathbb{E}[\mathbf{X}|\mathbf{Y}]$ evaluated with an LFPD results in a min-max estimator (i.e., an estimator that is best in the worst case). The case when conjugate priors are also LFPDs is viewed as desirable since it implies that the min-max estimators are also linear. Therefore, the characterization of the conjugate prior in Theorem 1 is an important step in this line of research. In this work, we do not seek to characterize the conditions under which the product gamma distribution is an LFPD. However, in [22] it has been demonstrated for the scalar Poisson case that the gamma distribution is an LPDF under variance and mean constraints (i.e., $\mathcal{P} = \{P_X : \mathbb{V}(X) \leq P, \mathbb{E}[X] \leq \mu\}$.

APPENDIX A
PROOF OF LEMMA 2

First, compute the cross-correlation term

$$\mathbb{E}[U_iU_j|\mathbf{Y}=\boldsymbol{y}] = \frac{\mathbb{E}[U_iU_jP_{\mathbf{Y}|\mathbf{U}}(\boldsymbol{y}|\mathbf{U})]}{P_{\mathbf{Y}}(\boldsymbol{y})} \tag{121}$$

$$= \frac{(y_i+1)(y_j+1)P_{\mathbf{Y}}(\boldsymbol{y}+\mathbf{1}_i+\mathbf{1}_j)}{P_{\mathbf{Y}}(\boldsymbol{y})}. \tag{122}$$

Therefore, by using Lemma 1

$$[\mathsf{Var}(\mathbf{U}|\mathbf{Y}=\boldsymbol{y})]_{ij}$$

$$= \frac{(y_i+1)(y_j+1)P_{\mathbf{Y}}(\boldsymbol{y}+\mathbf{1}_i+\mathbf{1}_j)}{P_{\mathbf{Y}}(\boldsymbol{y})} - \frac{(y_i+1)(y_j+1)P_{\mathbf{Y}}(\boldsymbol{y}+\mathbf{1}_i)P_{\mathbf{Y}}(\boldsymbol{y}+\mathbf{1}_j)}{P_{\mathbf{Y}}(\boldsymbol{y})P_{\mathbf{Y}}(\boldsymbol{y})} \tag{123}$$

$$=\mathbb{E}[U_i|\mathbf{Y}=\boldsymbol{y}]\left(\frac{(y_j+1)P_{\mathbf{Y}}(\boldsymbol{y}+\mathbf{1}_i+\mathbf{1}_j)}{P_{\mathbf{Y}}(\boldsymbol{y}+\mathbf{1}_i)} - \mathbb{E}[U_j|\mathbf{Y}=\boldsymbol{y}]\right) \tag{124}$$

$$= \mathbb{E}[U_i|\mathbf{Y}=\boldsymbol{y}]\left(\mathbb{E}[U_j|\mathbf{Y}=\boldsymbol{y}+\mathbf{1}_i] - \mathbb{E}[U_j|\mathbf{Y}=\boldsymbol{y}]\right). \tag{125}$$

This concludes the proof.

APPENDIX B
PROOF OF THEOREM 4

The proof for the Gaussian case is very similar to the Poisson case. We start with the following lemma.

Lemma 7: Let $\boldsymbol{\Sigma}$ be some covariance matrix and $\phi_{\mathbf{X}}(\boldsymbol{t})$ be the characteristic function of a random vector $\mathbf{X} \in \mathbb{R}^n$. Then, for every $\boldsymbol{t} \in \mathbb{R}^n$

$$\left|\phi_{\mathbf{X}}(\boldsymbol{t}) - \mathrm{e}^{-\frac{\boldsymbol{t}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{t}}{2}}\right| \le \|\boldsymbol{t}\| \max_{\tau\in[0,1]} \|\nabla\phi_{\mathbf{X}}(\tau\boldsymbol{t}) + \boldsymbol{\Sigma}\tau\boldsymbol{t}\phi_{\mathbf{X}}(\tau\boldsymbol{t})\|. \tag{126}$$

Proof: Let $\boldsymbol{r}(\tau) = \tau\boldsymbol{t}$

$$\left|\phi_{\mathbf{X}}(\boldsymbol{t})\,\mathrm{e}^{\frac{\boldsymbol{t}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{t}}{2}} - 1\right| \tag{127}$$

$$= \left|\int_0^1 \nabla\phi_{\mathbf{X}}(\boldsymbol{r}(\tau))\,\mathrm{e}^{\frac{\boldsymbol{r}(\tau)^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{r}(\tau)}{2}} \cdot \dot{\boldsymbol{r}}(\tau)\mathrm{d}\tau\right| \tag{128}$$

$$\le \int_0^1 \left|\nabla\phi_{\mathbf{X}}(\boldsymbol{r}(\tau))\,\mathrm{e}^{\frac{\boldsymbol{r}(\tau)^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{r}(\tau)}{2}} \cdot \dot{\boldsymbol{r}}(\tau)\right| \mathrm{d}\tau \tag{129}$$

$$\le \|\boldsymbol{t}\| \int_0^1 \mathrm{e}^{\tau^2\frac{\boldsymbol{t}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{t}}{2}} \|\nabla\phi_{\mathbf{X}}(\tau\boldsymbol{t}) + \boldsymbol{\Sigma}\tau\boldsymbol{t}\phi_{\mathbf{X}}(\tau\boldsymbol{t})\|\mathrm{d}\tau \tag{130}$$

$$\le \|\boldsymbol{t}\| \max_{\tau\in[0,1]} \|\nabla\phi_{\mathbf{X}}(\tau\boldsymbol{t}) + \tau\boldsymbol{\Sigma}\boldsymbol{t}\phi_{\mathbf{X}}(\tau\boldsymbol{t})\| \int_0^1 \mathrm{e}^{\tau^2\frac{\boldsymbol{t}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{t}}{2}}\mathrm{d}\tau \tag{131}$$

$$\le \|\boldsymbol{t}\| \max_{\tau\in[0,1]} \|\nabla\phi_{\mathbf{X}}(\tau\boldsymbol{t}) + \tau\boldsymbol{\Sigma}\boldsymbol{t}\phi_{\mathbf{X}}(\tau\boldsymbol{t})\|\,\mathrm{e}^{\frac{\boldsymbol{t}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{t}}{2}}, \tag{132}$$

where (128) follows from the fundamental theorem of calculus for line integrals; (129) follows from modulus inequality; and (130) is a consequence of using $\dot{\boldsymbol{r}}(\tau) = \boldsymbol{t}$, $\nabla\mathrm{e}^{\frac{\boldsymbol{t}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{t}}{2}} = \boldsymbol{\Sigma}\boldsymbol{t}\mathrm{e}^{\frac{\boldsymbol{t}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{t}}{2}}$ and the Cauchy–Schwarz inequality to produce the following sequence of bounds:

$$\left|\nabla\phi_{\mathbf{X}}(\boldsymbol{r}(\tau))\,\mathrm{e}^{\frac{\boldsymbol{r}(\tau)^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{r}(\tau)}{2}} \cdot \dot{\boldsymbol{r}}(\tau)\right|$$

$$= \mathrm{e}^{\frac{\tau^2\boldsymbol{t}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{t}}{2}} \left|(\nabla\phi_{\mathbf{X}}(\tau\boldsymbol{t}) + \tau\boldsymbol{\Sigma}\boldsymbol{t}\phi_{\mathbf{X}}(\tau\boldsymbol{t}))\cdot\boldsymbol{t}\right| \tag{133}$$

$$\le \mathrm{e}^{\frac{\tau^2\boldsymbol{t}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{t}}{2}} \|\nabla\phi_{\mathbf{X}}(\tau\boldsymbol{t}) + \tau\boldsymbol{\Sigma}\boldsymbol{t}\phi_{\mathbf{X}}(\tau\boldsymbol{t})\|\,\|\boldsymbol{t}\|. \tag{134}$$

This concludes the proof. ∎

Now, using the orthogonality principle, observe that

$$\mathbf{0} = \mathbb{E}\left[(\mathsf{A}\mathbf{X} - \mathbb{E}[\mathsf{A}\mathbf{X}|\mathbf{Y}])\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathbf{Y}}\right] \tag{135}$$

$$= \mathbb{E}\left[(\mathsf{A}\mathbf{X} - \mathsf{A}\mathsf{H}\mathbf{Y} + \mathsf{A}\mathsf{H}\mathbf{Y} - \mathbb{E}[\mathsf{A}\mathbf{X}|\mathbf{Y}])\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathbf{Y}}\right] \tag{136}$$

$$= \mathbb{E}\left[(\mathsf{A}\mathbf{X} - \mathsf{A}\mathsf{H}\mathbf{Y})\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathbf{Y}}\right] + \mathbb{E}\left[(\mathsf{A}\mathsf{H}\mathbf{Y} - \mathbb{E}[\mathsf{A}\mathbf{X}|\mathbf{Y}])\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathbf{Y}}\right]. \tag{137}$$

Moreover, the first term in (137) can be computed in terms of characteristic functions as follows:

$$\mathbb{E}\left[(\mathsf{A}\mathbf{X} - \mathsf{A}\mathsf{H}\mathbf{Y})\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathbf{Y}}\right]$$

$$= \mathbb{E}\left[(\mathsf{A}\mathbf{X} - \mathsf{A}\mathsf{H}\mathsf{A}\mathbf{X} - \mathsf{A}\mathsf{H}\mathbf{Z})\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathbf{Y}}\right] \tag{138}$$

$$= \mathbb{E}\left[(\mathsf{I} - \mathsf{A}\mathsf{H})\,\mathsf{A}\mathbf{X}\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathbf{Y}} - \mathsf{A}\mathsf{H}\mathbf{Z}\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathbf{Y}}\right] \tag{139}$$

$$= \mathbb{E}\left[(\mathsf{I} - \mathsf{A}\mathsf{H})\,\mathsf{A}\mathbf{X}\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathsf{A}\mathbf{X}}\right]\mathbb{E}\left[\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathbf{Z}}\right] \tag{140}$$

$$- \mathbb{E}\left[\mathsf{A}\mathsf{H}\mathbf{Z}\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathbf{Z}}\right]\mathbb{E}\left[\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathsf{A}\mathbf{X}}\right] \tag{141}$$

$$= (\mathsf{I} - \mathsf{A}\mathsf{H})\mathbb{E}\left[\mathsf{A}\mathbf{X}\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathsf{A}\mathbf{X}}\right]\phi_{\mathbf{Z}}(\boldsymbol{t}) - \mathsf{A}\mathsf{H}\mathbb{E}\left[\mathbf{Z}\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathbf{Z}}\right]\phi_{\mathsf{A}\mathbf{X}}(\boldsymbol{t}) \tag{142}$$

$$= (\mathsf{I} - \mathsf{A}\mathsf{H})\frac{1}{i}\nabla\phi_{\mathsf{A}\mathbf{X}}(\boldsymbol{t})\phi_{\mathbf{Z}}(\boldsymbol{t}) - \mathsf{A}\mathsf{H}\frac{1}{i}\nabla\phi_{\mathbf{Z}}(\boldsymbol{t})\phi_{\mathsf{A}\mathbf{X}}(\boldsymbol{t}) \tag{143}$$

$$= (\mathsf{I} - \mathsf{A}\mathsf{H})(-i)\nabla\phi_{\mathsf{A}\mathbf{X}}(\boldsymbol{t})\phi_{\mathbf{Z}}(\boldsymbol{t}) + \mathsf{A}\mathsf{H}\boldsymbol{t}(-i)\phi_{\mathbf{Z}}(\boldsymbol{t})\phi_{\mathsf{A}\mathbf{X}}(\boldsymbol{t}) \tag{144}$$

$$= (-i)\left((\mathsf{I} - \mathsf{A}\mathsf{H})\nabla\phi_{\mathsf{A}\mathbf{X}}(\boldsymbol{t}) + \mathsf{A}\mathsf{H}\boldsymbol{t}\phi_{\mathsf{A}\mathbf{X}}(\boldsymbol{t})\right)\phi_{\mathbf{Z}}(\boldsymbol{t}), \tag{145}$$

where (141) follows from the independence of $\mathbf{X}$ and $\mathbf{Z}$; (143) follows by observing that $\nabla\phi_{\mathsf{A}\mathbf{X}}(\boldsymbol{t}) = \mathbb{E}[i\mathsf{A}\mathbf{X}\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathsf{A}\mathbf{X}}]$ and $\nabla\phi_{\mathbf{Z}}(\boldsymbol{t}) = \mathbb{E}[i\mathbf{Z}\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathbf{Z}}]$; and (144) follows by using that $\nabla\phi_{\mathbf{Z}}(\boldsymbol{t}) = -\boldsymbol{t}\phi_{\mathbf{Z}}(\boldsymbol{t})$.

Next, by using (137) and (145), and applying the norm on both sides we get that

$$\left\|\mathbb{E}\left[\mathsf{A}(\mathsf{H}\mathbf{Y} - \mathbb{E}[\mathbf{X}|\mathbf{Y}])^\mathsf{T}\mathrm{e}^{i\boldsymbol{t}^\mathsf{T}\mathbf{Y}}\right]\right\|$$

$$= \|\phi_{\mathbf{Z}}(\boldsymbol{t})\left((\mathsf{I} - \mathsf{A}\mathsf{H})\nabla\phi_{\mathsf{A}\mathbf{X}}(\boldsymbol{t}) + \mathsf{A}\mathsf{H}\boldsymbol{t}\phi_{\mathsf{A}\mathbf{X}}(\boldsymbol{t})\right)\| \tag{146}$$

$$= \phi_{\mathbf{Z}}(\boldsymbol{t})\,\|(\mathsf{I} - \mathsf{A}\mathsf{H})\nabla\phi_{\mathsf{A}\mathbf{X}}(\boldsymbol{t}) + \mathsf{A}\mathsf{H}\boldsymbol{t}\phi_{\mathsf{A}\mathbf{X}}(\boldsymbol{t})\|. \tag{147}$$

Furthermore, by using the Cauchy–Schwarz inequality in (147)

$$\begin{aligned}
&\sqrt{\mathbb{E}\left[\|\mathbf{A}(\mathbf{H}\mathbf{Y}-\mathbb{E}[\mathbf{X}|\mathbf{Y}])\|^2\right]} \\
&\geq \phi_{\mathbf{Z}}(\boldsymbol{t})\left\|(\mathsf{I}-\mathbf{A}\mathbf{H})\nabla\phi_{\mathbf{A}\mathbf{X}}(\boldsymbol{t})+\mathbf{A}\mathbf{H}\boldsymbol{t}\phi_{\mathbf{A}\mathbf{X}}(\boldsymbol{t})\right\|, \qquad (148)\\
&\geq \phi_{\mathbf{Z}}(\boldsymbol{t})\sigma_{\min}(\mathsf{I}-\mathbf{A}\mathbf{H}) \\
&\quad\cdot\left\|\nabla\phi_{\mathbf{A}\mathbf{X}}(\boldsymbol{t})+(\mathsf{I}-\mathbf{A}\mathbf{H})^{-1}\mathbf{A}\mathbf{H}\boldsymbol{t}\phi_{\mathbf{A}\mathbf{X}}(\boldsymbol{t})\right\|, \qquad (149)
\end{aligned}$$

where (149) follows by using the fact that $(\mathsf{I}-\mathbf{A}\mathbf{H})$ is invertible and the inequality $\|\mathbf{A}\boldsymbol{x}\| \geq \sigma_{\min}(\mathbf{A})\|\boldsymbol{x}\|, \forall \boldsymbol{x}$ where $\sigma_{\min}(\mathbf{A})$ is the small singular value of $\mathbf{A}$.

Combining bounds in (126) and (149) and using the bound $\|\mathbf{A}\boldsymbol{x}\| \leq \|\mathbf{A}\|_{\star}\|\boldsymbol{x}\|, \forall \boldsymbol{x}$ we have that

$$\begin{aligned}
\frac{\left|\phi_{\mathbf{A}\mathbf{X}}(\boldsymbol{t})-\mathrm{e}^{-\frac{\boldsymbol{t}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{t}}{2}}\right|}{\|\boldsymbol{t}\|} &\leq \frac{\sqrt{\mathbb{E}\left[\|\mathbf{A}(\mathbf{H}\mathbf{Y}-\mathbb{E}[\mathbf{X}|\mathbf{Y}])\|^2\right]}}{\sigma_{\min}(\mathsf{I}-\mathbf{A}\mathbf{H})\,\phi_{\mathbf{Z}}(\boldsymbol{t})} \qquad (150)\\
&\leq \frac{\|\mathbf{A}\|_{\star}\sqrt{\mathbb{E}\left[\|\mathbf{H}\mathbf{Y}-\mathbb{E}[\mathbf{X}|\mathbf{Y}]\|^2\right]}}{\sigma_{\min}(\mathsf{I}-\mathbf{A}\mathbf{H})\,\phi_{\mathbf{Z}}(\boldsymbol{t})}, \qquad (151)
\end{aligned}$$

where $\boldsymbol{\Sigma}=(\mathsf{I}-\mathbf{A}\mathbf{H})^{-1}\mathbf{A}\mathbf{H}$. This concludes the proof.

REFERENCES

[1] A. Banerjee, X. Guo, and H. Wang, “On the optimality of conditional expectation as a Bregman predictor,” *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2664–2669, Jul. 2005.

[2] M. Raginsky, R. M. Willett, Z. T. Harmany, and R. F. Marcia, “Compressed sensing performance bounds under Poisson noise,” *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 3990–4002, Aug. 2010.

[3] L. Wang *et al.*, “Signal recovery and system calibration from multiple compressive Poisson measurements,” *SIAM J. Imag. Sci.*, vol. 8, no. 3, pp. 1923–1954, 2015.

[4] L. Wang, D. E. Carlson, M. Rodrigues, D. Wilcox, R. Calderbank, and L. Carin, “Designed measurements for vector count data,” in *Adv. Neural Inf. Process. Sys.*, 2013, pp. 1142–1150.

[5] S. Verdú, “Poisson communication theory,” in *Proc. Invited talk Int. Technion Commun. Day Honor Isr. Bar-David*, Mar. 1999.

[6] A. Lapidoth and S. M. Moser, “On the capacity of the discrete-time Poisson channel,” *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 303–322, Jan. 2009.

[7] D. Guo, S. Shamai, and S. Verdú, “Mutual information and conditional mean estimation in Poisson channels,” *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 1837–1849, May 2008.

[8] R. Atar and T. Weissman, “Mutual information, relative entropy, and estimation in the Poisson channel,” *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1302–1318, Mar. 2012.

[9] L. Wang, D. E. Carlson, M. R. Rodrigues, R. Calderbank, and L. Carin, “A Bregman matrix and the gradient of mutual information for vector Poisson and Gaussian channels,” *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2611–2629, May 2014.

[10] J. Grandell, *Mixed Poisson Processes*, Boca Raton, FL, USA: CRC Press, 1997, vol. 77.

[11] L. Wang and Y. Chi, “Stochastic approximation and memory-limited subspace tracking for Poisson streaming data,” *IEEE Trans. Signal Process.*, vol. 66, no. 4, pp. 1051–1064, Feb. 2018.

[12] A. Dytso and H. Vincent Poor, “Estimation in Poisson noise: Properties of the conditional mean estimator,” *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4304–4323, Jul. 2020.

[13] H. Raiffa and R. Schlaifer, *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Adminitration, Harvard University, 1961.

[14] P. Diaconis and D. Ylvisaker, “Conjugate priors for exponential families,” *Ann. Statist.*, vol. 7, no. 2, pp. 269–281, 1979.

[15] N. Johnson, “Uniqueness of a result in the theory of accident proneness,” *Biometrika*, vol. 44, no. 3–4, pp. 530–531, 1957.

[16] J.-P. Chou, “Characterization of conjugate priors for discrete exponential families,” *Statist. Sinica*, vol. 11, no. 2, pp. 409–418, 2001.

[17] I. J. Good, “The population frequencies of species and the estimation of population parameters,” *Biometrika*, vol. 40, no. 3–4, pp. 237–264, 1953.

[18] H. Robbins, “An empirical Bayes approach to statistics,” in *Proc. 3rd Berkeley Symp. Math Statist. Probab.*, Citeseer, 1956.

[19] D. P. Palomar and S. Verdú, “Representation of mutual information via input estimates,” *IEEE Trans. Inf. Theory*, vol. 53, no. 2, pp. 453–470, Feb. 2007.

[20] A. Dytso, R. Bustin, H. V. Poor, and S. S. Shitz, “On the equality condition for the I-MMSE proof of the entropy power inequality,” in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput.*, Monticello, IL, USA, Oct. 2017, pp. 1034–1039.

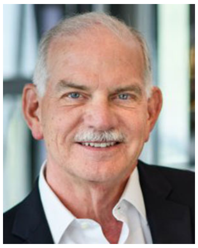
[21] F. du Pin Calmon, Y. Polyanskiy, and Y. Wu, “Strong data processing inequalities for input constrained additive noise channels,” *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1879–1892, 2018.

[22] A. Dytso, M. Fauß, and H. V. Poor, “A class of lower bounds for Bayesian risk with a Bregman loss,” in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun.*, Atlanta, GA, USA, 2020, pp. 1–5.

Alex Dytso (Member, IEEE) received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL, USA, in 2016. From September 2016 to August 2020, he was a Postdoctoral Associate with the Department of Electrical Engineering, Princeton University. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology (NJIT), Newark, NJ, USA. His current research interests are in the areas of multi-user information theory and estimation theory, and their applications in wireless networks.

Michael Fauß (Member, IEEE) received the Dipl.-Ing. degree from Technische Universität München, Munich, Germany, in 2010 and the Dr.-Ing. degree from Technische Universität Darmstadt, Darmstadt, Germany, in 2016, both in electrical engineering.

In November 2011, he joined the Signal Processing Group, Technische Universität Darmstadt, and in 2017 he received the dissertation award of the German Information Technology Society for his Ph.D. thesis on robust sequential detection. In September 2019, he joined Prof. H. Vincent Poor’s group with Princeton University as a Postdoctoral on a research grant by the German Research Foundation (DFG). His current research interests include statistical robustness, sequential detection and estimation, and the role of similarity measures in statistical inference.



H. Vincent Poor (Life Fellow, IEEE) received the Ph.D. degree in EECS from Princeton University, Princeton, NJ, USA, in 1977. From 1977 to 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990, he has been on the faculty at Princeton, where he is the Michael Henry Strater University Professor of Electrical Engineering. From 2006 to 2016, he served as the Dean of Princeton’s School of Engineering and Applied Science. He has also held visiting appointments at several other institutions, including most recently at Berkeley and Cambridge. His research interests are in the areas of information theory, signal processing and machine learning, and their applications in wireless networks, energy systems and related fields. Among his publications in these areas is the forthcoming book *Advanced Data Analytics for Power Systems* (Cambridge Univ. Press, 2021).

Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and is a foreign member of the Chinese Academy of Sciences, the Royal Society, and other national and international academies. He received the Technical Achievement and Society Awards of the IEEE Signal Processing Society in 2007 and 2011, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal, the 2019 ASEE Benjamin Garver Lamme Award, a D.Sc. *honoris causa* from Syracuse University, awarded in 2017, and a D.Eng. *honoris causa* from the University of Waterloo, awarded in 2019.