# Bayesian Risk With Bregman Loss: A Cramér–Rao Type Bound and Linear Estimation

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

***Abstract*— A general class of Bayesian lower bounds when the underlying loss function is a Bregman divergence is demonstrated. This class can be considered as an extension of the Weinstein–Weiss family of bounds for the mean squared error and relies on finding a variational characterization of Bayesian risk. This approach allows for the derivation of a version of the Cramér–Rao bound that is specific to a given Bregman divergence. This new generalization of the Cramér–Rao bound reduces to the classical one when the loss function is taken to be the Euclidean norm. In order to evaluate the effectiveness of the new lower bounds, the paper also develops upper bounds on Bayesian risk, which are based on optimal linear estimators. The effectiveness of the new bound is evaluated in the Poisson noise setting.**

***Index Terms*— Cramér–Rao, minimum mean squared error (MMSE), Bregman divergence, linear estimation, Poisson noise, Gaussian Noise.**

## I. Introduction

FINDING lower bounds on *a Bayesian risk* is an important issue in signal estimation as such bounds provide fundamental limits on signal recovery. Moreover, they can contribute useful insights and guidelines for algorithm design in data-driven applications, where Bayesian analysis is of ever-increasing importance. A plethora of such bounds are known for the mean squared error (MSE). Loosely speaking, these bounds can be divided into three families. The first family, termed *Weinstein–Weiss*, works by using the Cauchy–Schwarz inequality [2], and includes the prevalent *Cramér–Rao* (CR) bound (also known as the van-Trees bound [3]) as a special case. The second family, termed *Ziv–Zakai*, is derived by connecting estimation and binary hypothesis testing [4]. The third family uses a variational approach and works by minimizing the MSE subject to a constraint on a suitably chosen divergence measure, for example, the *Kullback–Leibler* (KL) divergence [5], [6].

This paper is concerned with a generalization of the Weinstein–Weiss family of bounds beyond the MSE. Specifically, the aim is to provide a generalization of this family to a larger class of Bayesian risks, where the loss functions are taken to be *Bregman divergences* (BDs). Bayesian risk based on BDs is playing an increasingly important role in estimation and information theory [7]–[11], and there is a need to derive lower bounds that will hold for these loss functions. A possible research program in this area consists of attempting to generalized each of the aforementioned family of bounds to the BD case. This work follows this program by generalizing the Weinstein–Weiss bounds, while follow-up work will generalize the variational bounds based on the Kullback-Leibler divergence. Beyond information theory, Bregman divergences are also starting to play an important role in statistics and machine learning, where non-Euclidean losses have found several applications [12]–[16]. The use of Bregman risk can also be motivated from the point of view of directional statistics [17].

The key to deriving the Weinstein–Weiss bounds for the MSE is the Cauchy–Schwarz inequality. The difficulty with such a generalization is that it is not immediately clear how the Cauchy–Schwarz inequality can be applied to BDs, which are, in general, not metrics, and do not necessarily have natural norms associated with them.

In this paper, by using elementary techniques such as Taylor's remainder theorem, it is shown that the Weinstein–Weiss approach can be generalized to the Bayesian risk when the loss function is taken to be a BD. Furthermore, this generalization makes it possible to derive a version of the CR bound that is specific to a given BD. This new generalization of the CR bound reduces to the classical CR bound when the loss function is taken to be the Euclidean norm.

In addition to developing a new Cramér–Rao type family of bounds, a second goal of this work is to assess the tightness of these bounds. However, in order to do so, we need to have access to the true Bayesian risk. The latter is usually not available, which is the reason why we seek lower bounds in the first place. Therefore, to assess the effectiveness of the lower bounds, we will further upper bound the Bayesian risk with a risk that uses a linear estimator and will compare the proximity of the upper and lower bounds. The theory of linear estimation is well-understood in the case of the MSE error. However, in the case of BDs, the general structure of optimal

Manuscript received February 3, 2021; revised November 11, 2021; accepted November 12, 2021. Date of publication November 23, 2021; date of current version February 17, 2022. 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. An earlier version of this paper was presented in part at the IEEE International Workshop on Signal Processing Advances in Wireless Communications [1] [DOI: 10.1109/SPAWC48557.2020.9154314]. *(Corresponding author: Alex Dytso.)*

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

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

Communicated by M. Lops, Associate Editor for Detection and Estimation.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2021.3130381.

Digital Object Identifier 10.1109/TIT.2021.3130381

linear estimators is not well-understood. To close this gap, this work will also focus on determining the optimal linear coefficients for a given BD.

The paper outline and contributions are as follows:

- Section II reviews properties of BDs and of the corresponding Bayesian risks.
- Section III presents a variational characterization of the Bregman risk and the new family of bounds.
- Section IV discusses some applications of the variational representation. In particular, Theorem 3 presents a generalization for the CR bound that is specific to a given BD and reduces to the classical CR bound when the Bayesian risk corresponds to the MSE.
- Section V discusses the structure of the optimal linear estimator for a given BD. In particular, Proposition 2 presents equations that characterize the optimal coefficients. Moreover, conditions on the prior distributions under which the optimal estimators are linear are discussed.
- Section VI, in order to show the utility of the new CR bound, evaluates the bound for the Poisson noise case with a BD natural for this setting. In particular, it is shown that the CR bound has the same behavior as the Bayesian risk when the scaling parameter of the Poisson noise is taken to be large.

*Notation:* Random variables are denoted by upper case letters, and their realizations are denoted by lower case letters. The inner product is denoted by $\langle \cdot, \cdot \rangle$. The identity matrix is denoted by $\mathsf{I}$. For two symmetric matrices $\mathsf{A}$ and $\mathsf{B}$ we say that $\mathsf{A} \prec \mathsf{B}$ if $\mathsf{B} - \mathsf{A}$ is positive-definite. For a symmetric positive semidefinite matrix $\mathsf{A}$ with an eigendecomposition $\mathsf{A} = \mathsf{V}\Lambda\mathsf{V}^{-1}$ the square root matrix is defined as $\mathsf{A}^{\frac{1}{2}} = \mathsf{V}\Lambda^{\frac{1}{2}}\mathsf{V}^{-1}$ where the square root of the diagonal matrix $\Lambda$ is defined element wise. For $0 \prec \mathsf{A}$ we define the Mahalanobis metric as $\|x\|_{\mathsf{A}} = \sqrt{x^T \mathsf{A} x}$, where $\|x\|$ denotes the Euclidian metric. The expected value is denoted by $\mathbb{E}[\cdot]$. For a random variable $X \in \mathbb{R}^n$ with a probability density function (pdf) $f_X$, the score function is defined as $\rho_X(x) = \frac{\nabla f_X(x)}{f_X(x)}$, where $\nabla$ is the gradient operator.

## II. Bregman Divergence and Bayesian Risk

In order to define a Bayesian risk or estimation error, one needs to select a loss function. The family of loss functions considered in this paper is defined next.

*Definition 1 (Bregman Divergence):* Let $\phi : \Omega \to \mathbb{R}$ be a *continuously-differentiable* and *strictly convex* function defined on a non-empty *closed convex* set $\Omega \subseteq \mathbb{R}^n \to [0, \infty)$. The Bregman divergence between $u$ and $v$ associated with the function $\phi$ is defined as

$$\ell_\phi(u, v) = \phi(u) - \phi(v) - \langle u - v, \nabla\phi(v) \rangle, \quad u, v \in \Omega. \quad (1)$$

*Remark 1:* Formally, to avoid issues with the differentiation of $\phi$, the Bregman divergence needs to be defined as $\ell_\phi : \Omega \times \mathrm{in}(\Omega) \to [0, \infty)$ where $\mathrm{in}(\Omega)$ denotes the relative interior of $\Omega$. However, for the ease of exposition we omit this notation.

The BD can be interpreted as an error due to an approximation of $\phi(u)$ with a line tangent to the point $(v, \phi(v))$. Fig. 1 illustrates this interpretation.

BDs have been introduced in [18] in the context of convex optimization. There exists several extension of the BD definition such as an extension to functional spaces [19], an extension to submodular set functions [20], and a matrix extension [21].

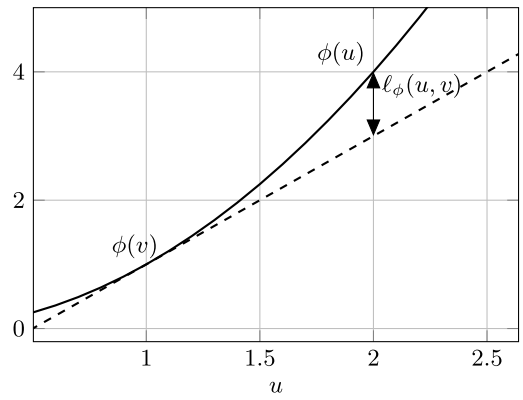



Fig. 1. Illustration of the definition of the BD.

In [22], BDs, together with $f$-divergences, were characterized axiomatically. A thorough investigation of BDs was undertaken in [12], where it was shown that many commonly used loss functions are members of this family. Moreover, the authors of [12] have shown that every regular exponential distribution has a unique BD associated with it.

Now consider the problem of estimating a random variable $X$ from a noisy observation $Y$, where the loss function is according to (1). The smallest Bayesian risk associated with this estimation problem is defined next.

*Definition 2 (Minimum Bayesian Risk With Respect to a BD):* For a joint distribution $P_{Y,X}$ we denote the minimum Bayesian risk with respect to the Bregman divergence $\ell_\phi(u, v)$ as

$$R_\phi(X|Y) = \inf_{f : f \text{ is measurable}} \mathbb{E}\left[\ell_\phi\left(X, f(Y)\right)\right]. \quad (2)$$

$R_\phi(X|Y)$ is also referred to as *Bayesian Bregman risk* in what follows.

*Remark 2:* The most prominent example of $R_\phi(X|Y)$ is the minimum mean squared error (MMSE), which is induced by choosing $\phi(u) = \|u\|^2$ and will be denoted by

$$\mathrm{mmse}(X|Y) = R_{\|\cdot\|^2}(X|Y). \quad (3)$$

The structure of the optimal estimator in (2) was studied in [23], where it was shown that the conditional expectation is the unique minimizer. Moreover, the authors of [23] have also demonstrated the converse result, namely that the conditional expectation is an optimal estimator only when the loss function is a Bregman divergence.

### A. Fundamental Properties of Bregman Divergences

We now, for completeness, review the most important properties of BDs and the associated Bayesian risks [12], [23].

*Theorem 1 (Fundamental Properties of Bregman Divergences and Bayesian Bregman Risks):*

1) *(Non-Negativity)* $\ell_\phi(u, v) \geq 0, \forall u, v \in \Omega$, with equality if and only if $u = v$;
2) *(Convexity)* $\ell_\phi(u, v)$ is convex in $u$;

3) *(Linearity)* $\ell_\phi(u,v)$ is linear in $\phi$;
4) *(Generalized Law of Cosines):* For $u,v,w\in\Omega$

$$\ell_\phi(u,v)=\ell_\phi(u,w)+\ell_\phi(w,v)-\langle u-w,\nabla\phi(v)-\nabla\phi(w)\rangle; \tag{4}$$

5) *(Orthogonality Principle and Pythagorean Identity)* For every random variable $X\in\Omega$ and every $u\in\Omega$

$$\mathbb{E}\left[\ell_\phi(X,u)\right]=\mathbb{E}\left[\ell_\phi(X,\mathbb{E}[X])\right]+\ell_\phi(\mathbb{E}[X],u). \tag{5}$$

Moreover, for any measurable $f(Y)$

$$\mathbb{E}\left[\ell_\phi(X,f(Y))\right]=\mathbb{E}\left[\ell_\phi\left(X,\mathbb{E}[X|Y]\right)\right]+\mathbb{E}\left[\ell_\phi\left(\mathbb{E}[X|Y],f(Y)\right)\right]. \tag{6}$$

6) *(Conditional Expectation is the Unique Minimizer)* Suppose that $\mathbb{E}[X]<\infty$ and $\mathbb{E}[\phi(X)]<\infty$. Then,

$$\inf_{f:f\text{ is measurable}}\mathbb{E}\left[\ell_\phi\left(X,f(Y)\right)\right]=\mathbb{E}\left[\ell_\phi\left(X,\mathbb{E}[X|Y]\right)\right]. \tag{7}$$

7) *(Coupling between Conditional Expectation and Bregman Divergence)* Let $F:\mathbb{R}^n\times\mathbb{R}^n\mapsto\mathbb{R}$ be a non-negative function such that $F(x,x)=0$ and assume that all partial derivatives $F_{x_i,x_j}$ are continuous. If for all random variables $X\in\mathbb{R}^n$ it holds that

$$\inf_{u\in\mathbb{R}^n}\mathbb{E}\left[F(X,u)\right]=\mathbb{E}\left[F(X,\mathbb{E}[X])\right], \tag{8}$$

with $\mathbb{E}[X]$ being the unique minimizer, then $F(u,v)=\ell_\phi(u,v)$ for some strictly convex and differentiable function $\phi:\mathbb{R}^n\to\mathbb{R}$.

### B. Notable Examples of BDs in Estimation Theory

BDs and their associated Bayesian risks appear naturally in connection with information measures. For example, the mutual information between $Y$ and $X$ can be represented as an integral of the BD induced by

- $\phi(u)=\|u\|^2$ (i.e., the MMSE) when $P_{Y|X}$ is a Gaussian distribution [24], [25];
- $\phi(u)=u\log u$ when $P_{Y|X}$ is a Poisson distribution [8], [9];
- $\phi(u)=u\log\frac{u}{1-u}$ when $P_{Y|X}$ is a binomial distribution [10]; and
- $\phi(u)=u\log\frac{u}{1+u}$ when $P_{Y|X}$ is a negative binomial distribution [10].

Table I summarizes the above examples together with the corresponding BDs. The latter will be referred to as the *natural* BDs in what follows. For a more detailed treatment of connections between information measures and BDs, the interested reader is referred to [26].

TABLE I
TABLE OF BDS

| Domain | $\phi(u)$ | $\ell_\phi(u,v)$ | Natural Noise |
|---|---|---|---|
| $\mathbb{R}^n$ | $\|u\|^2_{\mathsf{A}},\ \mathsf{0}\preceq\mathsf{A}$ | $\|u-v\|^2_{\mathsf{A}}$ | Gaussian |
| $\mathbb{R}_+$ | $u\log u$ | $u\log\frac{u}{v}-(u-v)$ | Poisson |
| $[0,1]$ | $u\log\frac{u}{1-u}$ | $u\log\frac{u(1-v)}{v(1-u)}-\frac{(u-v)}{1-v}$ | Binomial |
| $\mathbb{R}_+$ | $u\log\frac{u}{1+u}$ | $u\log\frac{u(1+v)}{v(1+u)}-\frac{(u-v)}{1+v}$ | Negative Binomial |

### A. Bregman Divergence Vs. Mahalanobis Distance

As one might expect, the variational characterization of the Bayesian Bregman risk requires an application of the Cauchy–Schwarz inequality. However, a priori, it is not immediately clear how the Cauchy–Schwarz inequality can be applied to frequently complicated expressions (see Table I) of BDs. The approach, however, becomes clear after an elementary application of Taylor's remainder theorem, which allows representing a BD as a weighted squared error.

*Lemma 1 ($\ell_2$-Representation of Bregman Divergences):* Suppose that $\phi$ in Definition 1 is twice differentiable and let

$$\Delta_\phi(u,v)=\frac{1}{2}\int_0^1(1-t)\mathsf{H}_\phi((1-t)u+tv)\mathrm{d}t, \tag{9}$$

where $x\mapsto\mathsf{H}_\phi(x)$ is the Hessian matrix of $\phi$ evaluated at $x$. Then, $\Delta_\phi(u,v)$ is positive definite and

$$\ell_\phi(u,v)=\|(u-v)^T\Delta_\phi^{\frac{1}{2}}(u,v)\|^2,\ u,v\in\Omega. \tag{10}$$

*Proof:* Recall that given a twice differentiable function $\phi:\mathbb{R}^n\to\mathbb{R}$ Taylor's remainder theorem [27] asserts that

$$\phi(u)=\phi(v)+\langle u-v,\nabla\phi(v)\rangle+\frac{1}{2}(u-v)^T\left[\int_0^1(1-t)\mathsf{H}_\phi(u+t(v-u))\mathrm{d}t\right](u-v). \tag{11}$$

Observe that the BD $\ell_\phi(u,v)$ is the remainder of the first order Taylor series expansion of $\phi(u)$ around $v$. Therefore, by the integral representation of the remainder in (11), it follows that

$$\ell_\phi(u,v)=\frac{1}{2}(u-v)^T\left[\int_0^1(1-t)\mathsf{H}_\phi(u+t(v-u))\mathrm{d}t\right](u-v) \tag{12}$$

$$=(u-v)^T\Delta_\phi(u,v)(u-v). \tag{13}$$

Furthermore, since $\phi$ is strictly convex, the Hessian matrix $\mathsf{H}_\phi$ and the matrix $\Delta_\phi(u,v)$ are positive definite. This concludes the proof. ■

*Remark 3:* In the scalar case (13) simplifies to

$$\ell_\phi(u,v)=\ell_{x^2}(u,v)\Delta_\phi(u,v), \tag{14}$$

with $\Delta_\phi$ being strictly positive.

*Remark 4:* It can be argued that (10) and (14) trivially hold true since any two functions with identical support can

## III. A VARIATIONAL REPRESENTATION OF BREGMAN RISK

In this section, we provide a variational characterization of Bayesian Bregman risk. We start by introducing an $\ell_2$-representation of Bregman divergences.

be expressed as weighted versions of one another by simply choosing the weight to be their ratio. Here, however, it is important to note that $\Delta_\phi$ can be obtained directly from $\phi$ by evaluating the integral on the right hand side of (9). In other words, $\Delta_\phi$ can be calculated without evaluating $\ell_\phi$.

We note there have been other attempts to connect $\ell_2$ distance and the BD. For instance, the authors of [28] also used the fact that BD is the remainder of the first order Taylor series expansion and expressed the remained as an infinite series. However, such infinite series representations require $\phi$ to be infinitely differentiable and do not lead to a compact representation as in (10).

Equation in (10) has a strong resemblance to Mahalanobis distance with the exception that the covariance matrix depends on the inputs $u, v$. That is, using (10), we can write

$$\ell_\phi(u,v) = \|u-v\|^2_{\Delta_\phi(u,v)}. \tag{15}$$

It is important to note that this analogy only holds *locally*, meaning that at any given point $(u, v)$ the BD $\ell_\phi(u, v)$ corresponds to a certain Mahalanobis distance. However, since the latter changes with $(u, v)$, the *global* properties of BDs and Mahalanobis distances, such as the expected risk considered here, can differ significantly. For instance, one can compare the topology of the ball induced by each divergence. To this end, let the Bregman-ball of radius $r$ with center at $c \in \Omega$ be defined as follows:

$$\mathcal{B}_\phi(r,c) = \{u \in \Omega : \ell_\phi(u,c) \le r\}. \tag{16}$$

Moreover, because $\ell_\phi(u, v)$ may not by symmetric we can also define $\tilde{\mathcal{B}}_\phi(r, c) = \{u \in \Omega : \ell_\phi(c, u) \le r\}$. For the Mahalanobis distance, the shape of the ball or the neighborhood is given by an ellipse. This is not the case for BDs. For example, consider the function $\phi(u) = u_1 \log u_1 + u_2 \log u_2$ where $u = [u_1, u_2]^T$ with $\Omega = \mathbb{R}^2_+$, which induces the following BD: for $u = [u_1, u_2]^T$ and $v = [v_1, v_2]^T$

$$\ell_\phi(u,v) = u_1 \log \frac{u_1}{v_1} + u_2 \log \frac{u_2}{v_2} - (u_1 - v_1) - (u_2 - v_2). \tag{17}$$

The BD in (17) is known as generalized I-divergence or generalized KL-divergence. Fig. 2 compares neighborhoods $\mathcal{B}_\phi(r, c)$ and $\tilde{\mathcal{B}}_\phi(r, c)$ induced by the BD in (17) to the standard Euclidian ball $\mathcal{B}_{\|x\|^2}(r, c)$ where we set $c = (2, 2)$ and $r = 1$.

### B. Variational Characterization of the Bayesian Bregman Risk

With Lemma 1 at our disposal, we are now ready to derive a variational characterization of the Bayesian Bregman risk.

*Theorem 2 (Variational Characterization of Bayesian Bregman Risk):* Let $g : \mathbb{R}^k \to \Omega$. Then,

$$\mathbb{E}\left[\ell_\phi(X, g(Y))\right] = \sup_{\psi:\mathbb{R}^n\times\mathbb{R}^k\to\mathbb{R}^n} \frac{\left|\mathbb{E}\left[(X-g(Y))^T\psi(X,Y)\right]\right|^2}{\mathbb{E}\left[\|\Delta_\phi^{-\frac{1}{2}}(X,g(Y))\psi(X,Y)\|^2\right]}, \tag{18}$$

and equality in (18) is attained if

$$\psi(X,Y) = \Delta_\phi(X, g(Y))(X - g(Y)). \tag{19}$$

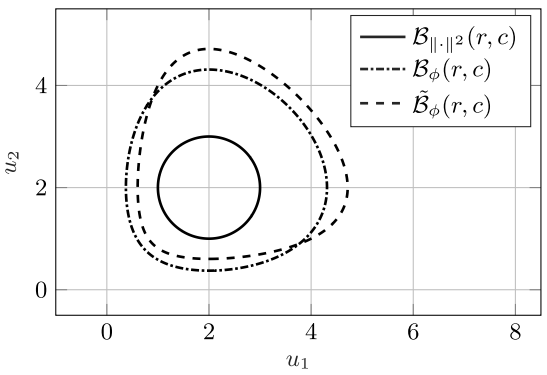



Fig. 2. Comparison of Bregman balls $\tilde{\mathcal{B}}_\phi(r, c), \tilde{\mathcal{B}}_\phi(r, c)$ and the Euclidean ball $\mathcal{B}_{\|x\|^2}(r, c)$ where $r = 1$ and $c = [2, 2]$.

*Proof:* By using Lemma 1 we have that

$$\mathbb{E}\left[\ell_\phi(X, g(Y))\right] = \mathbb{E}\left[\|(X-g(Y))^T\Delta_\phi^{\frac{1}{2}}(X, g(Y))\|^2\right] \tag{20}$$

$$\ge \frac{\left|\mathbb{E}\left[(X-g(Y))^T\Delta_\phi^{\frac{1}{2}}(X,g(Y))h(X,Y)\right]\right|^2}{\mathbb{E}\left[\|h(X,Y)\|^2\right]}, \tag{21}$$

where the last step follows from the Cauchy–Schwarz inequality for some arbitrary function $h : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$. Next, we rescale the expression by choosing $h(x, y) = \Delta_\phi^{-\frac{1}{2}}(x, g(y))\psi(x, y)$, for some arbitrary function $\psi : \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n$, which leads to the expression on right side of (18). The proof of the equality condition follows by inspection. This concludes the proof. ■

The variational characterization in (18) is a generalization of the Weinstein–Weiss representation of the MSE, which is included as the special case when $\Delta_\phi = \mathsf{I}$ [2]. Note that the expression in (18) holds even if the denominator on the right side of (18) vanishes but the numerator does not; in this case $\mathbb{E}\left[\ell_\phi(X, g(Y))\right] = \infty$.

Setting $g(Y) = \mathbb{E}[X|Y]$ yields a variational characterization of the minimum Bayesian risk with respect to a BD, which is an important corollary of the above result.

*Corollary 1:*

$$R_\phi(X|Y) = \sup_{\psi:\mathbb{R}^n\times\mathbb{R}^k\to\mathbb{R}^n} \frac{\left|\mathbb{E}\left[(X-\mathbb{E}[X|Y])^T\psi(X,Y)\right]\right|^2}{\mathbb{E}\left[\|\Delta_\phi^{-\frac{1}{2}}(X,\mathbb{E}[X|Y])\psi(X,Y)\|^2\right]}. \tag{22}$$

## IV. Discussion and Applications to Estimation

In this section, we first show a small application of the alternative representation of the BD in Lemma 1. Second, we present a generalized version of the CR that is specific to a given BD bound.

### A. Comparing Bregman Risk and the MMSE

Our first application shows how Bregman risks can be connected to the ubiquitous case of a risk with a squared error loss.

*Proposition 1:* Suppose that $\kappa_l \mathsf{I} \preceq \mathsf{H}_\phi \preceq \kappa_u \mathsf{I}$ for some constants $\kappa_l, \kappa_u \geq 0$. Then,

$$\kappa_l \mathrm{mmse}(X|Y) \leq R_\phi(X|Y) \leq \kappa_u \mathrm{mmse}(X|Y). \tag{23}$$

*Proof:* The proof follows by using Lemma 1, that is under the hypothesis of the theorem we have that

$$\kappa_l \mathsf{I} \preceq \Delta_\phi(u,v) \preceq \kappa_u \mathsf{I}. \tag{24}$$

Hence,

$$\kappa_l \ell_{x^2}(u,v) \leq \ell_\phi(u,v) \leq \kappa_u \ell_{x^2}(u,v). \tag{25}$$

This concludes the proof. ■

### B. A Generalization of the Bayesian CR Bound

The classic CR bound allows for lower bounding the MMSE with the Fisher information: for $X \in \Omega \subseteq \mathbb{R}^n$

$$\mathrm{mmse}(X|Y) \geq \frac{n^2}{\mathbb{E}\left[\|\nabla_X \log f_{YX}(Y,X)\|^2\right]}, \tag{26}$$

where the above holds under the regularity conditions

$$\mathbb{E}\left[\nabla_X \log f_{YX}(Y,X)|Y=y\right] = 0, \ \forall y; \ \text{and} \tag{27a}$$

$$x f_{YX}(y,x) = 0, \ \forall y, \ \forall x \in \partial\Omega, \tag{27b}$$

where $\partial\Omega$ denotes the boundary of the set $\Omega$ [3]. The quantity $\mathbb{E}\left[\|\nabla_X \log f_{YX}(Y,X)\|^2\right]$ is of course the Fisher information.

The next theorem proposes a generalization of the CR bound.

*Theorem 3 (Generalized CR-Bound):* Suppose that conditions in (27) hold. Then,

$$R_\phi(X|Y) \geq \frac{n^2}{\mathbb{E}\left[\|\Delta_\phi^{-\frac{1}{2}}(X,\mathbb{E}[X|Y])\nabla_X \log f_{YX}(Y,X)\|^2\right]}. \tag{28}$$

*Proof:* The proof follows by choosing $\psi(x,y) = \nabla_x \log f_{YX}(y,x)$ in (18). Now observe that

$$\begin{aligned}
&\mathbb{E}\left[(X-\mathbb{E}[X|Y])^T \nabla_X \log f_{YX}(Y,X)\right] \\
&= \mathbb{E}\left[X^T \nabla_X \log f_{YX}(Y,X)\right] \\
&\quad - \mathbb{E}\left[\mathbb{E}[X|Y]\mathbb{E}[\nabla_X \log f_{YX}(Y,X)|Y]\right] \qquad (29)\\
&= \mathbb{E}\left[X^T \nabla_X \log f_{YX}(Y,X)\right], \qquad (30)
\end{aligned}$$

where $\mathbb{E}\left[\mathbb{E}[X|Y]\mathbb{E}[\nabla_X \log f_{YX}(Y,X)|Y]\right] = 0$ from the assumption in (27). To conclude the proof note that

$$\begin{aligned}
&\mathbb{E}\left[X^T \nabla_X \log f_{YX}(Y,X)\right] \\
&= \int\int x^T \frac{\nabla_x f_{YX}(y,x)}{f_{YX}(y,x)} f_{YX}(y,x)\mathrm{d}x\mathrm{d}y \qquad (31)\\
&= \int\int x^T \nabla_x f_{YX}(y,x)\mathrm{d}x\mathrm{d}y \qquad (32)\\
&= \sum_{i=1}^n \int\int x_i \frac{\partial}{\partial x_i} f_{YX}(y,x)\mathrm{d}x\mathrm{d}y = -n, \qquad (33)
\end{aligned}$$

where in the last step we have used integration by parts and the fact that $x f_{YX}(y,x) = 0$ for $x \in \partial\Omega$. ■

Observe that the quantity in (28) can be thought of as a generalization of the Fisher information that takes into account the corresponding BD. Its dependence on the quantity $\Delta_\phi^{-\frac{1}{2}}(X,\mathbb{E}[X|Y])$ makes the generalized CR bound more difficult to compute than the classical CR bound for the MMSE for which $\Delta_\phi^{-\frac{1}{2}}(X,\mathbb{E}[X|Y]) = \mathsf{I}$. In Section VI, by using a Poisson noise example, we will show how this difficulty can be overcome. We also refer the reader to [29] where we show yet another example in which the lower bound in (28) is effective for the binomial noise channel. A goal of this work is to assess the tightness of the above bound. However, in order to make such a comparison, we need to have access to $R_\phi(X|Y)$, which is not usually available and is the reason why we find lower bounds in the first place. Therefore, in order to assess the effectiveness of the lower bound, we will first upper bound $R_\phi(X|Y)$ and then compare the proximity of the upper and lower bounds. Our upper bound on $R_\phi(X|Y)$ will be based on a risk that uses an optimal *linear* estimator instead of the optimal estimator. The theory of linear estimation is well-understood in the case of the MSE error but is less well understood in the case of BD divergences. In the next section, we present several results concerning optimal linear estimators for the BDs.

*Remark 5:* We note that (28) is not the only way of generalizing the CR bound. There exist several other generalizations of the CR bound either to other loss functions or other notions of variance; the interested reader is referred to [30]–[37] and the references therein.

*Remark 6:* An interesting feature of the CR lower bound in (28) is that it depends on the estimator $g(Y)$. Note that in the case of the MSE the CR bound in (26) does not depend on the estimator in question and is uniform over all estimators. On the one hand, a benefit of this dependence is that one can adapt the lower bound to the estimator in use and potentially get a tighter bound. On the other hand, a drawback might be the computability of such a bound. However, the latter can be addressed by using the CR bound corresponding to $R_\phi(X|Y)$, which is uniform over all estimators, i.e.,

$$\begin{aligned}
\mathbb{E}\left[\ell_\phi(X,g(Y))\right] &\geq R_\phi(X|Y) \\
&\geq \frac{n^2}{\mathbb{E}\left[\|\Delta_\phi^{-\frac{1}{2}}(X,\mathbb{E}[X|Y])\nabla_X \log f_{YX}(Y,X)\|^2\right]}.
\end{aligned}$$

Finally, we would like to note that in the non-Bayesian literature lower bounds on risk often depend on the estimator. An example of such a bound is the CR bound for biased estimators; see for example [38].

## V. On Optimal Linear Estimators

We say that a linear estimator $g(Y) = \mathsf{C}Y + d$, where $\mathsf{C} \in \mathbb{R}^{n\times k}$ and $d \in \mathbb{R}^n$, is permissible with respect to the domain $\Omega \subseteq \mathbb{R}^n$ (or simply permissible) if

$$\mathbb{P}[\mathsf{C}Y + d \in \Omega] = 1. \tag{34}$$

Note that unlike for the MSE where $\Omega = \mathbb{R}^n$, not all linear estimators might be permissible for a given BD as

$g(Y) = \mathsf{C}Y + d$ might not belong to $\Omega$. Consequently, we define the minimum linear Bayesian risk as

$$R_{\phi,\mathsf{L}}(X|Y) = \inf_{g: g=\mathsf{C}Y+d, \text{ and g is permissible}} \mathbb{E}\left[\ell_\phi(X; g(Y))\right]. \tag{35}$$

It immediately follows that $R_\phi(X|Y) \leq R_{\phi,\mathsf{L}}(X|Y)$. We next explore the following two questions: 1) What is a characterization of the optimal coefficients in (35)? and 2) When are linear estimators optimal (i.e., $R_\phi(X|Y) = R_{\phi,\mathsf{L}}(X|Y)$)?

### A. On the Characterization of Optimal Linear Estimators

Recall that the optimal coefficients of MSE-optimal linear estimators are given by [39]

$$\mathsf{C}_{\text{mse}} = \mathsf{Cov}(X,Y)\mathsf{Var}^{-1}(Y), \tag{36a}$$
$$d_{\text{mse}} = \mathbb{E}[X] - \mathsf{C}^\star\mathbb{E}[Y], \tag{36b}$$

where $\mathsf{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])^T] \in \mathbb{R}^{n\times k}$ and $\mathsf{Var}(Y) = \mathbb{E}[(Y-\mathbb{E}[Y])(Y-\mathbb{E}[Y])^T] \in \mathbb{R}^{k\times k}$. It is important to note that the MSE-optimal coefficients are not in general optimal for other BDs. The next result provides necessary conditions for the optimality of linear estimators.

*Proposition 2:* Fix some $\phi$. Then, $\mathsf{C}, d$ minimize (35) only if the following conditions hold:

$$\mathbb{P}[\mathsf{C}Y + d \in \Omega] = 1, \tag{37a}$$
$$\mathbb{E}[\mathsf{H}_\phi(\mathsf{C}Y+d)(\mathsf{C}Y+d-X)Y^T] = 0, \tag{37b}$$
$$\mathbb{E}[\mathsf{H}_\phi(\mathsf{C}Y+d)(\mathsf{C}Y+d-X)] = 0. \tag{37c}$$

The above conditions are also sufficient if $v \to \ell_\phi(u,v)$ is convex for very $u \in \Omega$.

*Proof:* The condition in (37a) is needed for the estimator to be permissible with respect to the domain $\Omega$. To show the other two conditions let $\tilde{\mathsf{C}} = [\mathsf{C}\,|\,d]$ and $\tilde{Y}^T = [Y^T\ 1]$. We now check the first order condition necessary for optimality

$$0 = \nabla_{\tilde{\mathsf{C}}}\mathbb{E}[\ell_\phi(X, \tilde{\mathsf{C}}\tilde{Y})]. \tag{38}$$

If $v \to \ell_\phi(u,v)$ is convex for very $u \in \Omega$, then the first order condition is also sufficient.

To find the gradient first observe that

$$\nabla_v \ell_\phi(u,v) = \mathsf{H}_\phi(v)(v-u). \tag{39}$$

Consequently, by using the chain-rule of differentiation we have that

$$\nabla_{\tilde{\mathsf{C}}}\ell_\phi(u, \tilde{\mathsf{C}}\tilde{y}) = \nabla\ell_\phi(u, \tilde{\mathsf{C}}\tilde{y})\tilde{y}^T \tag{40}$$
$$= \mathsf{H}_\phi(\tilde{\mathsf{C}}v)(\tilde{\mathsf{C}}\tilde{y} - u)\tilde{y}^T. \tag{41}$$

Therefore, the first order condition becomes

$$0 = \nabla_{\tilde{\mathsf{C}}}\mathbb{E}[\ell_\phi(X, \tilde{\mathsf{C}}\tilde{Y})] = \mathbb{E}[\mathsf{H}_\phi(\tilde{\mathsf{C}}\tilde{Y})(\tilde{\mathsf{C}}\tilde{Y} - X)\tilde{Y}^T]. \tag{42}$$

■

Perhaps the most obvious example to consider first is that of Gaussian statistics.

*Example:* Fix some $\phi$ with $\Omega = \mathbb{R}^n$ and $(X,Y)$ let be jointly Gaussian. Then, the Bregman-optimal linear estimator for $R_{\phi,\mathsf{L}}(X|Y)$ is given by the MSE-optimal estimator in (36b).

*Proof:* First, we show that for Gaussian statistics, the coefficients of the estimator in (36b) solve the equations in (37c), which show that these estimators satisfy the necessary condition. The fact that these are also sufficient will be shown in a more general result in the next section.

Since $\Omega = \mathbb{R}^n$, we only need to check the last two equations in (37c). These simplify to

$$0 = \mathbb{E}\left[\mathbb{E}\left[\mathsf{H}_\phi(\mathsf{C}Y+d)(\mathsf{C}Y+d-X)Y^T|Y\right]\right]$$
$$= \mathbb{E}\left[\mathsf{H}_\phi(\mathsf{C}Y+d)(\mathsf{C}Y+d-\mathbb{E}[X|Y])Y^T\right], \text{ and} \tag{43}$$
$$0 = \mathbb{E}\left[\mathbb{E}\left[\mathsf{H}_\phi(\mathsf{C}Y+d)(\mathsf{C}Y+d-X)|Y\right]\right]$$
$$= \mathbb{E}\left[\mathsf{H}_\phi(\mathsf{C}Y+d)(\mathsf{C}Y+d-\mathbb{E}[X|Y])\right]. \tag{44}$$

The fact that the above equations are equal to zero if $(\mathsf{C}, d)$ are chosen to be the coefficients of the MSE-optimal estimator follows from the fact that if $(X,Y)$ are jointly Gaussian, then $\mathbb{E}[X|Y]$ is given by the linear estimator in (36b). ■

*Remark 7:* It is interesting to note that, under Gaussian statistics, the MSE-optimal linear estimator is optimal for a much wider range of loss functions such as $\ell_p, p \geq 1$. Moreover, for these loss functions, the conditional expectation is, in general, not an optimal estimator. The interested reader is referred to [40] and references therein.

*Remark 8:* In general, the equations in (37c) can be difficult to compute in closed-form, and one needs to resort to algorithmic solutions. We do not attempt to explore this question in depth as it warrants its own independent study. However, one possible approach to finding optimal coefficients is to use a *projected gradient descent algorithm*. To that end, let $\tilde{\mathsf{C}} = [\mathsf{C}\,|\,d]$ and $\tilde{Y}^T = [Y^T\ 1]$ and

$$\mathcal{C} = \{\tilde{\mathsf{C}} : \mathbb{P}[\tilde{\mathsf{C}}\tilde{Y} \in \Omega] = 1\}. \tag{45}$$

The set $\mathcal{C}$ represents the collection of all permissible linear estimators with respect to $\Omega$. The projection operation onto $\mathcal{C}$ is denoted by $\mathsf{proj}_{\mathcal{C}}(\cdot)$. Therefore, the update equation for the gradient descent is given by

$$\tilde{\mathsf{C}}_{t+1} = \mathsf{proj}_{\mathcal{C}}\left(\tilde{\mathsf{C}}_t - \lambda\nabla_{\tilde{\mathsf{C}}}\mathbb{E}[\ell_\phi(X, \tilde{\mathsf{C}}\tilde{Y})]\right),\ t \in \mathbb{N} \tag{46}$$

where $\lambda \geq 0$ and where the gradient was computed in (42) and is given by $\nabla_{\tilde{\mathsf{C}}}\mathbb{E}[\ell_\phi(X, \tilde{\mathsf{C}}\tilde{Y})] = \mathbb{E}[\mathsf{H}_\phi(\tilde{\mathsf{C}}\tilde{Y})(\tilde{\mathsf{C}}\tilde{Y} - X)\tilde{Y}^T]$.

The difference between the MSE-optimal coefficients $(\mathsf{C}_{\text{mse}}, d_{\text{mse}})$ and the Bregman-optimal coefficients $(\mathsf{C}^\star, d^\star)$ will be demonstrated via several examples in Section VI in the context of Poisson noise. A partial answer to when $(\mathsf{C}_{\text{mse}}, d_{\text{mse}}) = (\mathsf{C}^\star, d^\star)$ is given next.

### B. Conjugate Priors and Exponential Family

In this section, we consider the question: when does $R_{\phi,\mathsf{L}}(X|Y) = R_\phi(X|Y)$? Clearly, a sufficient condition is equality between the optimal estimator $\mathbb{E}[X|Y]$ and the optimal linear estimators. This condition turns out to be also necessary, as is shown next.

*Lemma 2:* Let the linear estimator $g(Y) = \mathsf{C}^\star Y + d^\star$ be Bregman-optimal in the sense of (35). Then,

$$R_{\phi,\mathsf{L}}(X|Y) = R_\phi(X|Y), \tag{47}$$

if and only if

$$\mathbb{E}[X|Y] = \mathsf{C}^\star Y + d^\star, \;\; Y\text{-almost surely.} \tag{48}$$

Moreover, $(\mathsf{C}^\star, d^\star)$ are the MSE-optimal coefficients given in (36b).

*Proof:* Using the Pythagorean property in Theorem 1 we have that

$$R_{\phi,\mathsf{L}}(X|Y) - R_\phi(X|Y) = \mathbb{E}[\ell_\phi(\mathbb{E}[X|Y], \mathsf{C}^\star Y + d^\star)]. \tag{49}$$

Therefore, $R_{\phi,\mathsf{L}}(X|Y) = R_\phi(X|Y)$ if and only if $\mathbb{E}[\ell_\phi(\mathbb{E}[X|Y], \mathsf{C}^\star Y + d^\star)] = 0$. Since BDs are non-negative, this is true if and only if $\ell_\phi(\mathbb{E}[X|Y], \mathsf{C}^\star Y + d^\star) = 0$, $Y$-almost surely. Moreover, in view of the first property in Theorem 1, $\ell_\phi(\mathbb{E}[X|Y], \mathsf{C}^\star Y + d^\star) = 0$, $Y$-almost surely, if and only if $\mathbb{E}[X|Y] = \mathsf{C}^\star Y + d^\star$, $Y$-almost surely.

Finally, because $\mathbb{E}[X|Y]$ is also MSE-optimal, the estimator $\mathsf{C}^\star Y + d^\star$ must also be the MSE-optimal linear estimator. ■

In view of Lemma 2, the condition for the equality of $R_{\phi,\mathsf{L}}(X|Y)$ and $R_\phi(X|Y)$ reduces to characterizing a set of distributions on $(X, Y)$ such that (48) holds. Answering this question in full generality is a difficult task. However, partial answers to this question are available in the literature. In particular, a very general set of distributions is known for the case when

$$\mathbb{E}[X|Y] = a^\star Y + d^\star, \tag{50}$$

that is when $\mathsf{C}^\star = a^\star \mathsf{I}$ for some scalar $a^\star$. We next present this family of distributions. This is done by first defining a set of noise transformations $P_{Y|X}$, and then defining the set of priors $P_X$.

The noise transformation that we use comes from the exponential family defined next [41].

*Definition 3:* An $n$-parameter natural regular exponential family $\{P_{Y|\Theta=\theta}\}_y$ of probability measures with respect to a $\sigma$-finite measure $\mu$ on $\mathbb{R}^n$ is given by

$$\frac{\mathrm{d}P_{Y|\Theta=\theta}}{\mathrm{d}\mu}(y) \;= \mathrm{e}^{\langle y,\theta\rangle - \psi(\theta)}, \theta \in \mathcal{N}, \tag{51}$$

where $y \in \mathbb{R}^n$ and $\mathcal{N} \subset \mathbb{R}^n$ is an open set, and where

$$\mathrm{e}^{\psi(\theta)} = \int \mathrm{e}^{\langle y,\theta\rangle} \mathrm{d}\mu(y). \tag{52}$$

The parameter $\Theta$, the set $\mathcal{N}$ and the function $\psi$ are known as the natural parameter, the natural parameter space, and the log-partition function, respectively.

We denote the mean parameter of the exponential family by $X = \mathbb{E}[Y|\Theta]$. Recall that there is a one-to-one mapping between the natural parameter and the mean parameter:

$$X = \nabla\psi(\Theta) \text{ and } \Theta = \nabla\psi^*(X), \tag{53}$$

where $\psi^*$ is a convex conjugate of $\psi$.[1]

As an object of estimation, we will focus on the random mean parameter $X$ with a distribution $P_X$, and let the random transformation according to (51) be denoted by

$$Y = \mathcal{E}_{\psi,\mu}(X). \tag{54}$$

[1]Let $f : \Omega \to \mathbb{R}$ be a convex function on $\Omega$. Then, its *convex conjugate* is the function $f^*(y) = \sup_{x\in\Omega}(\langle y, x\rangle - f(x)), y \in \Omega$.

We will also refer to the expression in (54) as a channel. We next define a set of priors on $X$ for which linear estimation is optimal. Such priors are easiest to define with respect to the natural parameter $\Theta$.

*Definition 4 (Conjugate Prior):* Let $\mathcal{N}$ be a nonempty convex open set in $\mathbb{R}^n$ and let $\lambda$ denote the Lebesgue measure. Define a measure $\tilde{\Pi}_{\phi,\eta,\nu} \ll \lambda$ whose density with respect to $\lambda$ is given by

$$\tilde{\pi}_\phi(\theta|\eta,\nu) = \mathrm{e}^{\langle \eta,\theta\rangle - \nu\psi(\theta)}, \; \nu \in \mathbb{R}, \; \eta \in \mathbb{R}^n, \theta \in \mathcal{N}, \tag{55}$$

where $\psi(\theta)$ is a log-partition function. If $\tilde{\Pi}_{\psi,\eta,\nu}(\mathcal{N}) < \infty$, then $\tilde{\pi}_\psi(\theta|\eta,\nu)$ can be normalized to a probability measure with density $\pi_\psi(\theta|\eta,\nu)$. We refer to $\pi_\psi(\theta|\eta,\nu)$ as the conjugate prior associated with $\psi$.

*Remark 9:* Let $\mathcal{Y} \subseteq \mathbb{R}^n$ denote the interior of the convex hull of the support set of the measure $\mu$ in Definition 3. As was show in [42], a sufficient and necessary condition for $\tilde{\Pi}_{\eta,\nu}(\mathcal{N}) < \infty$ is that $\nu > 0$ and $\frac{\eta}{\upsilon} \in \mathcal{Y}$. In the remaining analysis we assume that this condition holds.

*Theorem 4:* Assume the following:

- $\nabla\psi(\mathcal{N}) \subseteq \Omega$; and
- $Y = \mathcal{E}_{\psi,\mu}(X)$ and $\Theta \sim \pi_\psi(\cdot|\eta,\nu)$.

Then,

$$R_\phi(X|Y) = R_{\phi,\mathsf{L}}(X|Y), \tag{56}$$

where $\mathbb{E}[X|Y = y] = \frac{\eta+y}{\nu+1}$ (i.e., $a^\star = \frac{1}{\eta+1} \in \mathbb{R}$ and $d^\star = \frac{\eta}{\upsilon+1} \in \mathbb{R}^n$).

*Proof:* The assumption that $\Theta$ follows a conjugate prior $\pi_\psi(\cdot|\eta,\nu)$ together with the result in [42, Thm. 2] imply that

$$\mathbb{E}[\nabla\phi(\Theta)|Y = y] = \frac{\eta + y}{\nu + 1}. \tag{57}$$

Moreover, using the fact that $X = \nabla\phi(\Theta)$, we have that $\mathbb{E}[X|Y = y] = \frac{\eta+y}{\nu+1}$. Furthermore, the condition $\nabla\psi(\mathcal{N}) \subseteq \Omega$ implies that $\mathbb{E}[X|Y = y] = \frac{\eta+y}{\nu+1}$ and $X$ are permissible with respect to the domain $\Omega$, where $X$ is permissible in an almost sure sense. Hence, both are valid input arguments into the Bergman divergence $\ell_\phi$.

Now since the conditional expectation is a linear function of $Y$, the conclusion that $R_\phi(X|Y) = R_{\phi,\mathsf{L}}(X|Y)$ follows from Lemma 2. ■

*Remark 10:* Despite the fact that for conjugate priors we have a simpler structure for the conditional expectation, finding a closed-form expression for $R_{\phi,\mathsf{L}}(X|Y)$ can still be a difficult task.

Further investigation of the conditions under which the conditional expectation is a linear function is an interesting director for further research. For example, one ambitious direction is to characterize the set of pairs of priors and exponential distributions for which $\mathbb{E}[X|Y] = \mathsf{C}Y + d$ where $\mathsf{C}$ is not an identity matrix. As was shown in [43] for the vector Poisson channel, such pairs do not always exist, and the only case when $\mathbb{E}[X|Y]$ is linear is when $\mathsf{C}$ is a diagonal matrix.

## VI. Evaluations of the Bounds for the Poisson Noise Case

In this section, we consider an observation model governed by Poisson noise and a loss function natural for this setting.

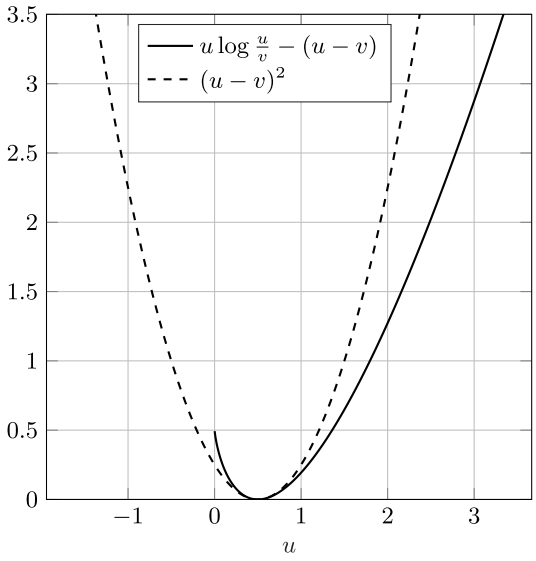



Fig. 3. Comparison of the squared error loss to the loss in (58) for $v = 0.5$.

Specifically, we consider $\phi(u) = u \log u$ with $\Omega = \mathbb{R}_+$ so that

$$\ell_\phi(u; v) = u \log \frac{u}{v} - (u - v),\ u, v \in \Omega \tag{58}$$

which is natural for Poisson noise. Note that $\phi''(u)$ is unbounded, and the results of Proposition 1 do not apply. Therefore, it is non-trivial to compare the Bayesian risk corresponding to (58) and the MMSE. Fig. 3 compares the squared error loss to the loss in (58).

Next, consider the problem of denoising a non-negative random variable in *Poisson noise*. The random Poisson transformation of a non-negative real-valued input random variable $X$ to a non-negative integer-valued output random variable $Y$ will be denoted by

$$Y = \mathcal{P}(aX), \tag{59}$$

where $a > 0$ is the *scaling factor*. Concretely, the Poisson noise channel is dictated by the following probability mass function (pmf):

$$P_{Y|X}(y|x) = \frac{1}{y!}(ax)^y \mathrm{e}^{-(ax)}, \tag{60}$$

where $y = 0, 1, \ldots$ and $x \geq 0$. In words, conditioned on a non-negative input $X$, the output of the Poisson channel is a non-negative integer-valued random variable $Y$ that is distributed according to (60). Note that in (60) we use the convention that $0^0 = 1$. Poisson noise models comprise an important family of models with a wide range of applications, including optical communications [44], [45]. The analysis of the MMSE in the context of Poisson noise was undertaken in [46]–[48]. Our objectives in this section are the following:

1) Compare Bregman-optimal coefficients for linear estimation $(c^\star, d^\star)$ and MSE-optimal coefficients for linear estimation $(c_{\text{mse}}, d_{\text{mse}})$. Moreover, we also want to compare $R_{\phi,\mathsf{L}}(X|Y)$ and $\mathbb{E}\left[\ell_\phi(X; c_{\text{mse}}Y + d_{\text{mse}})\right]$. In particular, it will be shown that $(c^\star, d^\star)$ may not be equal to $(c_{\text{mse}}, d_{\text{mse}})$. Moreover, it will be shown that the linear estimator with coefficients $(c^\star, d^\star)$ may be biased, while the linear MMSE estimator is always unbiased.
2) Evaluate the effectiveness of the CR bound. In particular, it will be shown that the CR bound is effective in the high signal-to-noise-ratio regime; and
3) Understand the scaling of $R_\phi(X|Y)$ and $\mathrm{mmse}(X|Y)$ as a function of the parameter $a$.

Note that to achieve 2) and 3), we will need to find a prior on $X$ such that $R_{\phi,\mathsf{L}}(X|Y), R_\phi(X|Y)$ and $\mathrm{mmse}(X|Y)$ can be computed exactly or very efficiently numerically. To find such a distribution, we resort to the notion of conjugate priors for which $R_{\phi,\mathsf{L}}(X|Y) = R_\phi(X|Y)$ discussed in Section V-B.

### A. On Optimal Linear Estimation

The MSE-optimal linear estimator in (36b) has a closed-form expression that depends only on the second-order statistics and can be easily implemented. Therefore, an interesting question to explore is how does the MSE-optimal linear estimator in (36b) perform for a given BD when compared to the Bregman-optimal linear estimator? In other words, we seek to compare $R_{\phi,\mathsf{L}}(X|Y)$ and $\mathbb{E}\left[\ell_\phi(X; c_{\text{mse}}Y + d_{\text{mse}})\right]$.

We begin by finding the best linear estimator for Poisson noise and the corresponding MSE. Besides helping us answer the aforementioned question, this computation will also help us to evaluate the effectiveness of the CR bound in Section VI-D.

*Lemma 3 (Best Linear Estimator for the MSE):* Let $Y = \mathcal{P}(aX)$, $\phi(u) = u^2$. Then,

$$c_{\text{mse}} = \frac{\mathbb{V}(X)}{a\mathbb{V}(X) + \mathbb{E}[X]},\ d_{\text{mse}} = \frac{\mathbb{E}^2[X]}{a\mathbb{V}(X) + \mathbb{E}[X]}, \tag{61}$$

and

$$R_{\phi,\mathsf{L}}(X|Y) = \frac{\mathbb{V}(X)}{a\frac{\mathbb{V}(X)}{\mathbb{E}[X]} + 1}. \tag{62}$$

*Proof:* See Appendix A. ■

For the loss function induced by $\phi(u) = u \log(u)$, we have that $v \to \ell_\phi(u; v)$ is convex for every $u$, and the conditions in Proposition 2 become sufficient and necessary and reduce to

$$\mathbb{E}\left[\frac{XY}{cY + d}\right] = a\mathbb{E}[X], \tag{63a}$$

$$\mathbb{E}\left[\frac{X}{cY + d}\right] = 1, \tag{63b}$$

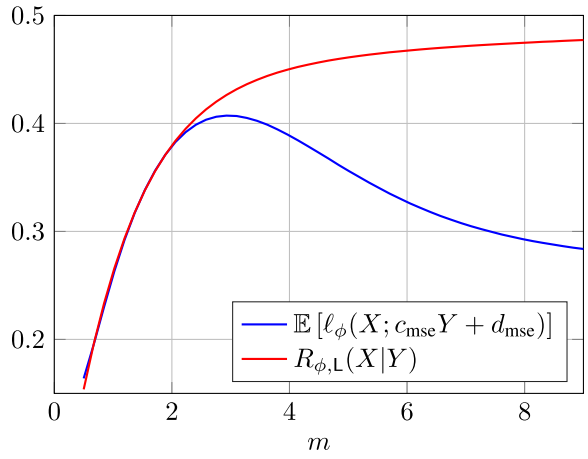
where $c \geq 0, d \geq 0$. As before the coefficients that solve the above equations are denoted by $(c^\star, d^\star)$. In general, the equations in (63b) do not appear to have a closed-form solution and must be solved numerically.

We now consider an example that shows that the Bregman-optimal coefficients can differ from the MSE-optimal coefficients. Consider the case when the input random variable $X$ is distributed equally likely on $\{0, m\}$ (i.e., on-off signaling) and let $Y = \mathcal{P}(X)$. Then, using (61) we have that

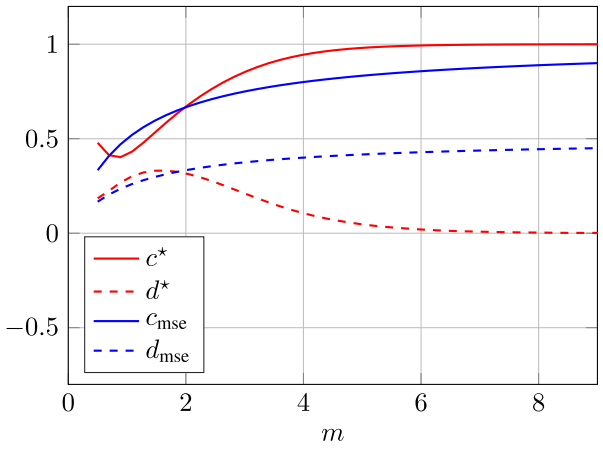
$$c_{\text{mse}} = \frac{m}{m+1},\quad d_{\text{mse}} = \frac{m/2}{m+1}. \tag{64}$$

Moreover, for $\phi(u) = u \log u$

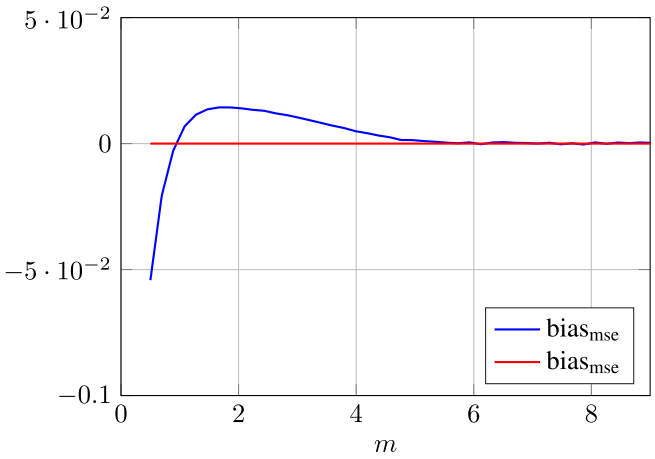
$$\mathbb{E}\left[\ell_\phi(X; c_{\text{mse}}Y + d_{\text{mse}})\right] = \frac{m}{2}\log(m+1) - \frac{m}{2}\mathbb{E}\left[\log\left(N_m + \frac{1}{2}\right)\right], \tag{65}$$




(a) $R_{\phi,\mathsf{L}}(X|Y)$ vs. $\mathbb{E}\left[\ell_\phi(X; c_{\text{mse}}Y + d_{\text{mse}})\right]$.




(b) $(c^\star, d^\star)$ vs. $(c_{\text{mse}}, d_{\text{mse}})$.




(c) Bias comparison where $\text{bias}_{\text{mse}} = \mathbb{E}[X - c_{\text{mse}}Y + d_{\text{mse}}]$ and $\text{bias} = \mathbb{E}[X - c^\star Y + d^\star]$.

Fig. 4. Comparison of $R_{\phi,\mathsf{L}}(X|Y)$ and $\mathbb{E}\left[\ell_\phi(X; c_{\text{mse}}Y + d_{\text{mse}})\right]$.

where $N_m$ is a Poisson random variable with mean $m$. This type of on-off signaling is commonly used over the Poisson noise channel, for example, as a modulation scheme [45].

We now use the algorithm in Remark 8 to explore the question of how does the mismatched linear risk $\mathbb{E}\left[\ell_\phi(X; c_{\text{mse}}Y + d_{\text{mse}})\right]$ compare to the Bregman-optimal linear risk $R_{\phi,\mathsf{L}}(X|Y)$. Fig. 4 depicts the following: 1) the performance of $R_{\phi,\mathsf{L}}(X|Y)$ vs. the performance of $\mathbb{E}\left[\ell_\phi(X; c_{\text{mse}}Y + d_{\text{mse}})\right]$; 2) the Bregman-optimal coefficients $(c^\star, d^\star)$ for $R_{\phi,\mathsf{L}}(X|Y)$ and the MSE-optimal coefficients $(c_{\text{mse}}, d_{\text{mse}})$; and 3) the bias of the best linear estimator with respect to $R_{\phi,\mathsf{L}}(X|Y)$.

From the above example we observe the following:

1) The values of $\mathbb{E}\left[\ell_\phi(X; c_{\text{mse}}Y + d_{\text{mse}})\right]$ and $R_{\phi,\mathsf{L}}(X|Y)$ can differ significantly. See Fig. 4a for the comparison.
2) The values of the MSE-optimal coefficients $(c_{\text{mse}}, d_{\text{mse}})$ and the Bregman-optimal coefficients $(c^\star, d^\star)$ for $R_{\phi,\mathsf{L}}(X|Y)$ can different significantly. See Fig. 4b for the comparison.
3) While the linear MMSE estimator is always unbiased, the best linear Bregman estimator can be biased. See Fig. 4c for an example.

In the above example, we have set the scaling coefficient to $a = 1$ and looked at the behavior of the risk as a function of the distribution of $X$ by increasing the parameter $m$. We now fix the distribution of $X$ and vary the scaling parameter $a$. By inserting the MSE coefficients into the optimality equation in (63b), it is not difficult to check that

$$\lim_{a\to 0}(c^\star, d^\star) = \lim_{a\to 0}(c_{\text{mse}}, d_{\text{mse}}),$$
$$\lim_{a\to \infty}(c^\star, d^\star) = \lim_{a\to \infty}(c_{\text{mse}}, d_{\text{mse}}).$$

In other words, for a fixed distribution, the MSE-optimal coefficients are Bregman-optimal both in high and low noise regimes.

### *B. The Conjugate Prior for Poisson Noise*

In order to study the estimation in Poisson noise, it is useful to consider the conjugate prior distribution for this noise. In the case of Poisson noise, we can describe the conjugate prior directly in terms $X$. A conjugate prior for the Poisson distribution is given by a gamma distribution. We say that $X$ is distributed according to a gamma distribution if it has a pdf given by

$$f(x) = \frac{\alpha^\theta}{\Gamma(\theta)} x^{\theta-1} \mathrm{e}^{-\alpha x},\ x \geq 0, \tag{66}$$

where $\theta > 0$ is the shape parameter and $\alpha > 0$ is the rate parameter. We denote the distribution with the pdf in (66) by $\mathsf{Gam}(\alpha, \theta)$.

The next lemma compiles several properties of the gamma distribution needed in this section.

*Lemma 4:* Suppose that $X \sim \mathsf{Gam}(\alpha, \theta)$ and that $Y = \mathcal{P}(aX)$. Then,

$$\mathbb{E}[X] = \frac{\theta}{\alpha}, \tag{67}$$

$$\mathbb{V}(X) = \frac{\theta}{\alpha^2}, \tag{68}$$

$$\mathbb{E}\left[\frac{1}{X^n}\right] = \begin{cases} \infty, & \theta \leq n \\ \alpha^n \frac{\Gamma(\theta-n)}{\Gamma(\theta)}, & \theta > n \end{cases}, \tag{69}$$

$$\mathbb{E}[\rho_X^2(X)] = \begin{cases} \infty, & \theta \leq 2 \\ \frac{\alpha^2}{\theta-2} & \theta > 2 \end{cases}, \tag{70}$$

$$\mathbb{E}[\rho_X^2(X)X] = \begin{cases} \infty, & \theta \leq 1 \\ \alpha, & \theta > 1 \end{cases}, \tag{71}$$

$$\mathbb{E}[X|Y=y] = \frac{1}{\alpha+a}y + \frac{\theta}{\alpha+a}. \tag{72}$$

*Proof:* See Appendix B. ∎

### C. Regularity Conditions

We now verify conditions under which the CR bound in (28) holds. This condition is given in (27) and is independent of the choice of $\phi$. It is also important to observe that the output of the Poisson noise channel is discrete. However, there is no issue in applying the CR bound in (28) as differentiability is only required in the $X$ variable, while the $Y$ variable can have an arbitrary support.

*Proposition 3 (CR Regularity Condition for Poisson Noise):* Let $X \sim f_X$ and $Y = \mathcal{P}(aX)$. The conditions in (27) hold if

$$\lim_{x \to 0^+} f_X(x) = 0. \tag{73}$$

*Proof:* See Appendix C. ∎

We now applying the above condition to our conjugate prior.

*Lemma 5:* Let $X \sim \mathsf{Gam}(\alpha, \theta)$. Then, the regularity condition in (73) holds if $\theta > 1$.

*Proof:* The fact that the regularity condition in (73) holds for $\theta > 1$ follows from the limit

$$\lim_{x \to 0^+} x^{\theta-1}\mathrm{e}^{-\alpha x} = \begin{cases} \infty & \theta < 1 \\ 1, & \theta = 1 \\ 0, & \theta > 1 \end{cases}. \tag{74}$$

∎

### D. CR Bound for the Squared Loss

In this section, we compute the CR bound for the MMSE. This computation has two purposes. First, we will be able to compare the performance of the MMSE to that of $R_\phi(X|Y)$. Second, this computation is of independent interest as the MSE is still a widely use fidelity criterion, and, to the best of our knowledge, the Bayesian CR bound has not been computed in the Poisson noise case.

*Theorem 5 (CR Bound for the MMSE):* Let $Y = \mathcal{P}(aX)$ and suppose (73) holds. Then,

$$\mathrm{mmse}(X|Y) \geq \frac{1}{a\mathbb{E}\left[\frac{1}{X}\right] + \mathbb{E}\left[\rho_X^2(X)\right]}. \tag{75}$$

*Proof:* See Appendix D. ∎

From Lemma 3 we have an upper bound on $\mathrm{mmse}(X|Y)$, and from Theorem 5 we have a lower bound on $\mathrm{mmse}(X|Y)$ from which we conclude that

$$\mathrm{mmse}(X|Y) = \Theta\left(\frac{1}{a}\right). \tag{76}$$

In other words, the CR bound is an effective lower bound for large values of $a$, which corresponds to the low noise regime. Further evidence of the effectiveness of the bound will be given in Section VI-F when the bound will be evaluated with a gamma distribution.

### E. CR Bound for the Natural BD

Before computing the CR bound in Theorem 3 for $\phi(u) = u\log u$, we present two ancillary lemmas. The first result provides bounds on $\Delta_\phi(u, v)$.

*Lemma 6:* Let $\phi(u) = u\log u$. Then,

$$2u \leq \frac{1}{\Delta_\phi(u,v)} \leq \frac{8u}{3} + \frac{4v}{3}. \tag{77}$$

*Proof:* See Appendix E. ∎

*Lemma 7:* Let $Y = \mathcal{P}(aX)$ with $a > 0$. Then,

$$\lim_{y \to \infty} \frac{\mathbb{E}[X|Y=y]}{(y+1)} \leq \frac{1}{a}. \tag{78}$$

Consequently, there exist constants $c_1$ and $c_2$ such that

$$\mathbb{E}[X|Y=y] \leq c_1 y + c_2, \quad y = 0, 1, \ldots \tag{79}$$

Furthermore, $c_1$ and $c_2$ can be choose as

$$c_1 = c_2 = \sup_{n \geq 0} \frac{\mathcal{M}_X^{(n+1)}(-a)}{(n+1)\mathcal{M}_X^{(n)}(-a)}, \tag{80}$$

where $\mathcal{M}_X(t) = \mathbb{E}\left[\mathrm{e}^{tX}\right], t \geq 0$ and $\mathcal{M}_X^{(n)}(t)$ is the moment generating function of $X$ and its $n$-th derivative.

*Proof:* See Appendix F. ∎

The result in Lemma 7 says that we can find a linear estimator that upper bounds $\mathbb{E}[X|Y]$. This further underscores the importance of linear estimators.

We now proceed to evaluate the CR lower bound in Theorem 3.

*Theorem 6 (CR Bound):* Let $\phi(u) = u\log u$ and $Y = \mathcal{P}(aX)$ and suppose that regularity condition in (73) holds. Then,

$$R_\phi(X|Y) \geq \frac{1}{D_{c_1,c_2}}, \tag{81}$$

where $c_1$ and $c_2$ are defined in (79) and where

$$\begin{aligned} D_{c_1,c_2} = {} & \frac{8}{3}\left(a + \mathbb{E}\left[\rho_X^2(X)X\right]\right) \\ & + \frac{4c_1}{3}\left(a^2 + a\mathbb{E}\left[\frac{1}{X}\right] + a\mathbb{E}\left[\rho_X^2(X)X\right]\right) \\ & + \frac{4c_2}{3}\left(a\mathbb{E}\left[\frac{1}{X}\right] + \mathbb{E}\left[\rho_X^2(X)\right]\right). \end{aligned} \tag{82}$$

*Proof:* See Appendix G. ∎

Note that the lower bound on the MMSE in (75) depends on $E\left[\frac{1}{X}\right]$ and $\mathbb{E}\left[\rho_X^2(X)\right]$ while the lower bound in (81) additionally depends on the product term $\mathbb{E}\left[\rho_X^2(X)X\right]$. In what

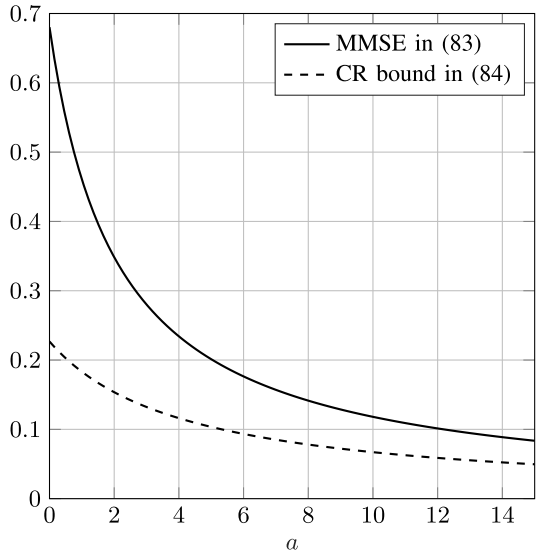



Fig. 5. Evaluation of the exact MMSE in (83) and the CR bound in (84). The parameters are set to $\alpha = 2.1$ and $\theta = 3$.

follows, we further evaluate the bounds in (75) and (81) with the gamma prior.

### *F. Evaluations of the CR Bounds With a Gamma Prior*

In this section, we evaluate the CR bound for the MMSE and the Bregman Risk induced by the function $\phi(u) = u \log u$. In addition, to assess the tightness of these bounds, we provide either closed-form or easily computable expressions for the risk. Note that, in general, we do not have the luxury of closed-form expressions, and the case with the gamma prior is one of the few cases for which it is possible to obtain such expressions.

The next result evaluates the MMSE and the CR lower bound for the case when $X \sim \mathsf{Gam}(\alpha, \theta)$.

*Proposition 4 (Gamma Prior and MMSE):* Suppose that $X \sim \mathsf{Gam}(\alpha, \theta)$ and $Y = \mathcal{P}(aX)$. Then, the following statements hold:

- (An Exact Expression) An exact expression for the MMSE is given by

$$\mathrm{mmse}(X|Y) = \frac{\mathbb{V}(X)}{a\frac{\mathbb{V}(X)}{\mathbb{E}[X]} + 1} = \frac{\theta}{\alpha(a+\alpha)}, \tag{83}$$

  with $\mathbb{V}(X) = \frac{\theta}{\alpha^2}$ and $\mathbb{E}(X) = \frac{\theta}{\alpha}$;
- (CR Bound) The CR regularity condition in (73) holds for $\theta > 1$. Moreover, the bound in (75) reduces to

$$\mathrm{mmse}(X|Y) \geq \begin{cases} 0, & 1 < \theta \leq 2 \\ \frac{\theta-1}{\alpha\left(a+\alpha\frac{\theta-1}{\theta-2}\right)}, & \theta \geq 2 \end{cases}. \tag{84}$$

*Proof:* See Appendix H. ■

Fig. 5 compares the exact MMSE and the CR lower bound evaluated in Proposition 4.

*Remark 11:* The fact that the bound in (62) is attained by a gamma distribution suggest that a gamma distribution is *the least favorable prior* distribution for the MMSE under the mean and variance constraint. To put it differently, for $Y = \mathcal{P}(aX)$ the maximizer of

$$P_{X^\star} = \arg \max_{P_X : \mathbb{E}[X]=\mu, \mathbb{V}(X)=P} \mathrm{mmse}(X|Y), \tag{85}$$

is given by $P_{X^\star} = \mathsf{Gam}(\alpha^\star, \theta^\star)$ with $\alpha^\star = \left(\frac{\mu}{P}\right)^{\frac{1}{4}}$ and $\theta^\star = \mu^{\frac{3}{4}} P^{\frac{1}{4}}$. We also note that the maximizer in (85) is unique; see [49, Remark 9] for the details.

As noted in Remark 10, the fact that the estimator has a simple form does not necessarily imply that $R_\phi(X|Y)$ does too. Therefore, in order to evaluate the tightness of the CR bound, we also provide an exact expression for $R_\phi(X|Y)$ and an upper bound on $R_\phi(X|Y)$, both of which are amenable to numerical evaluation.

*Proposition 5 (Gamma Prior for BD):* Let $\phi(u) = u \log u$, $Y = \mathcal{P}(aX)$ and $X \sim \mathsf{Gam}(\alpha, \theta)$. Then, the following statements hold:

- (CR Bound) The constant $c_1$ and $c_2$ in the CR bound in Theorem 6 can be chosen to be $c_1 = c^\star, c_2 = d^\star$ and

$$D_{c^\star, d^\star} = \frac{8}{3}(a+\alpha) + \frac{4}{3}\frac{1}{\alpha+a}\left(a^2 + a\frac{\alpha}{\theta-1} + a\alpha\right) + \frac{4}{3}\frac{\theta}{\alpha+a}\left(a\frac{\alpha}{\theta-1} + \frac{\alpha^2}{\theta-2}\right), \tag{86}$$

  for $\theta > 2$ and $D_{c^\star, d^\star} = \infty$ for $\theta \in [1, 2]$.
- (An Exact Expression)

$$R_\phi(X|Y) = \mathbb{E}[X \log X] - B, \tag{87}$$

  where

$$\mathbb{E}[X \log X] = \frac{\theta\left(\log\left(\frac{1}{\alpha}\right) + \psi(\theta+1)\right)}{\alpha}, \tag{88}$$

  where $\psi$ is the digamma function, and

$$B = \mathbb{E}\left[\left(\frac{Y}{\alpha+a} + \frac{\theta}{\alpha+a}\right)\log\left(\frac{Y}{\alpha+a} + \frac{\theta}{\alpha+a}\right)\right], \tag{89}$$

  and where $Y$ has a negative binomial distribution with the pmf given by

$$P_Y(y) = \frac{a^y \alpha^\theta}{(\alpha+a)^{\theta+y}}\binom{\theta+y-1}{y},\ y = 0, 1, \ldots \tag{90}$$

- (An Upper on $R_\phi(X|Y)$) The expression in (89) can be further lower bounded as follows:

$$B \geq \mathbb{E}\left[\left(\frac{aX}{\alpha+a} + \frac{\theta}{\alpha+a}\right)\log\left(\frac{aX}{\alpha+a} + \frac{\theta}{\alpha+a}\right)\right]. \tag{91}$$

*Proof:* See Appendix I. ■

Fig. 6 compares the CR bound in (86) to the exact value and bounds on $R_\phi(X|Y)$ computed in Proposition 5.

*Remark 12:* Note that in Proposition 5 the CR bound in (86) can be computed analytically while the exact expression for $R_\phi(X|Y)$ in (89) cannot be computed in closed-form and needs to be evaluated numerically. One possible way to do this is to use the Monte-Carlo approach and generate samples of $Y$ according to the pmf in (90). Alternatively, since the random variable $Y$ is discrete with countable support, one can compute (89) by truncating the infinite sum after sufficiently many

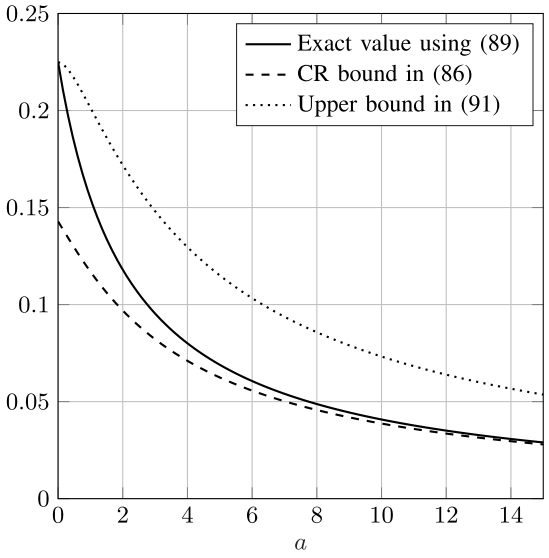



Fig. 6. Evaluation of the exact Bregman risk in (89), the CR bound in (86), and the bounds in (91). The parameters are set to $\alpha = 2.1$ and $\theta = 3$.

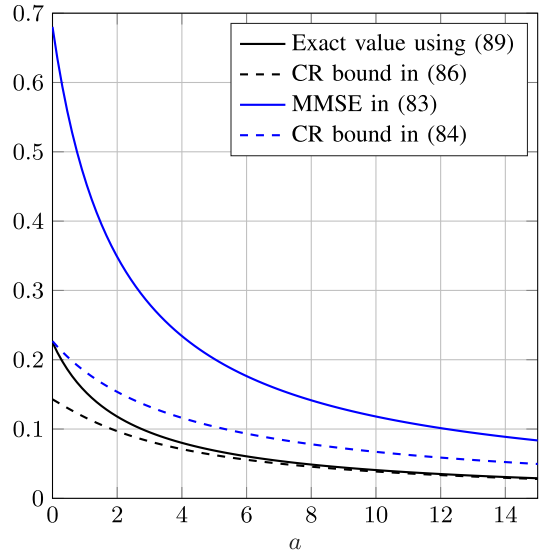



Fig. 7. Scaling of the MMSE and the Bregman risk with $\phi(u) = u \log u$. The parameters are set to $\alpha = 2.1$ and $\theta = 3$.

terms. To compute $R_\phi(X|Y)$, we take the latter approach and truncate the sum in (89) at $y = 300$.

It is also instructive to loosen the lower bound in (91) to

$$B \geq \mathbb{E}\left[\left(\frac{aX}{\alpha + a}\right) \log\left(\frac{aX}{\alpha + a}\right)\right], \tag{92}$$

which implies that

$$R_\phi(X|Y) \leq \frac{\mathbb{E}\left[X \log X\right]}{\alpha + a} - \frac{(a\mathbb{E}\left[X\right] + \theta) \log\left(\frac{a}{\alpha + a}\right)}{\alpha + a}. \tag{93}$$

Therefore, the new CR bound is of the same order as the upper bound and

$$R_\phi(X|Y) = \Theta\left(\frac{1}{a}\right). \tag{94}$$

This conclusion demonstrates that the new CR bound is effective.

Finally, note that both $R_\phi(X|Y)$ in (94) and the MMSE in (83) are of the same order. Fig. 7 compares the scaling of the MMSE and the Bregman risk induced by $\phi(u) = u \log u$ for the gamma prior. From Fig. 7 we also see that the scaling of the MMSE and $R_\phi(X|Y)$ do not match as $a \to 0$. Indeed, it is not difficult to show the following limits:

$$\lim_{a \to 0} R_\phi(X|Y) = \mathbb{E}[X \log X] - \mathbb{E}[X] \log(\mathbb{E}[X]),$$

$$\lim_{a \to 0} \mathrm{mmse}(X|Y) = \mathbb{V}(X).$$

Note that this behavior of the CR bound is not unexpected, and it is well-known that the CR bound is good only in the high signal-to-noise-ratio regime [3], while typically different types of lower bounds are needed in the low signal-to-noise-ratio regime.

## VII. Conclusion

This paper has proposed a general class of Bayesian lower bounds for the case in which the underlying loss function is a BD. The approach allows for deriving a version of the CR bound that is specific to a given BD. To show the applicability of the new CR bound it has been evaluated for the Poisson noise case. For both examples, the bounds have been shown to admit the same behavior as the corresponding Bregman risk in the low noise regime, hence, demonstrating the effectiveness of the new CR bound.

## Appendix A
## Proof of Lemma 3

The minimizers of (61) are given in (36b), and the minimum value is given by

$$R_{\phi,\mathrm{L}}(X|Y) = \mathbb{V}(X) - (c^\star)^2 \mathbb{V}(Y). \tag{95}$$

The proof is completed by observing that

$$\mathbb{E}[Y] = \mathbb{E}[aX] = a\mathbb{E}[X], \tag{96}$$

$$\mathbb{V}(Y) = a^2 \mathbb{V}(X) + a\mathbb{E}[X], \tag{97}$$

$$\mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] = a\mathbb{V}(X). \tag{98}$$

## Appendix B
## Proof of Lemma 4

The expressions for $\mathbb{E}[X]$ and $\mathbb{V}(X)$ are standard. To show (69) we use the following well-known integral:

$$\int_0^\infty x^k \mathrm{e}^{-\alpha x} \mathrm{d}x = \begin{cases} \infty, & k \leq -1, \\ \frac{\Gamma(k+1)}{\alpha^{k+1}}, & k > -1. \end{cases} \tag{99}$$

Therefore,

$$\mathbb{E}\left[\frac{1}{X^n}\right] = \frac{\alpha^\theta}{\Gamma(\theta)} \int x^{\theta - 1 - n} \mathrm{e}^{-\alpha} \mathrm{d}x = \alpha^n \frac{\Gamma(\theta - n)}{\Gamma(\theta)}, \tag{100}$$

for $\theta > n$ and infinity otherwise.

To show (99) observe that

$$\rho_X(x) = \frac{\theta - 1}{x} - \alpha,\ x > 0 \tag{101}$$

and, hence,

$$\mathbb{E}\left[\left(\frac{\theta-1}{X}-\alpha\right)^2\right]=\mathbb{E}\left[\frac{(\theta-1)^2}{X^2}-2\alpha\frac{\theta-1}{X}+\alpha^2\right] \quad (102)$$

$$=\begin{cases}\infty, & \theta\le 2,\\ \frac{(\theta-1)\alpha^2}{(\theta-2)}-2\alpha^2+\alpha^2, & \theta>2,\end{cases} \quad (103)$$

where we have used that $\mathbb{E}\left[\frac{1}{X^2}\right]=\alpha^2\frac{\Gamma(\theta-2)}{\Gamma(\theta)}=\frac{\alpha^2}{(\theta-2)(\theta-1)}$ for $\theta>2$ and $\mathbb{E}\left[\frac{1}{X}\right]=\alpha\frac{\Gamma(\theta-1)}{\Gamma(\theta)}=\frac{\alpha}{\theta-1}$ for $\theta>1$. Moreover,

$$\mathbb{E}\left[\left(\frac{\theta-1}{X}-\alpha\right)^2 X\right]=\mathbb{E}\left[\frac{(\theta-1)^2}{X}-2\alpha(\theta-1)+\alpha^2 X\right]$$

$$=\begin{cases}\infty, & \theta\le 1\\ \alpha, & \theta>1\end{cases}, \quad (104)$$

where have used that $\mathbb{E}\left[\frac{1}{X}\right]=\alpha\frac{\Gamma(\theta-1)}{\Gamma(\theta)}=\frac{\alpha}{\theta-1}$ for $\theta>1$.

The proof of (72) can be found in [49]. This concludes the proof.

## Appendix C
## Proof of Proposition 3

$$\mathbb{E}\left[\nabla_X \log\left(P_{Y|X}(Y|X)f_X(X)\right)|Y=y\right]$$

$$=\int_0^\infty \frac{\nabla_x\left(P_{Y|X}(y|x)f_X(x)\right)}{P_{Y|X}(y|x)f_X(x)}f_{X|Y}(x|y)\mathrm{d}x \quad (105)$$

$$=\frac{1}{P_Y(y)}\int_0^\infty \nabla_x\left(P_{Y|X}(y|x)f_X(x)\right)\mathrm{d}x \quad (106)$$

$$=\frac{1}{P_Y(y)}\left(P_{Y|X}(y|x)f_X(x)\Big|_0^\infty\right) \quad (107)$$

$$=-\frac{1}{P_Y(y)}\lim_{x\to 0^+}P_{Y|X}(y|x)f_X(x)=0, \quad (108)$$

where the last step follows from the assumption in (73). This verifies that the CR bound applies.

## Appendix D
## Proof of Theorem 5

Observe that the score function is readily computed to be

$$\nabla_x \log\left(P_{Y|X}(y|x)f_X(x)\right)=\frac{\nabla_x P_{Y|X}(y|x)}{P_{Y|X}(y|x)}+\frac{\nabla_x f_X(x)}{f_X(x)}$$

$$=\frac{y}{x}-a+\rho_X(x). \quad (109)$$

Therefore, the denominator in the CR bound is given by

$$\mathbb{E}\left[\left(\frac{Y}{X}-a+\rho_X(X)\right)^2\right]$$

$$=\mathbb{E}\left[\left(\frac{Y}{X}-a\right)^2\right]+2\mathbb{E}\left[\left(\frac{Y}{X}-a\right)\rho_X(X)\right]+\mathbb{E}\left[\rho_X^2(X)\right]$$

$$=\mathbb{E}\left[\left(\frac{Y}{X}-a\right)^2\right]+\mathbb{E}\left[\rho_X^2(X)\right] \quad (110)$$

$$=a\mathbb{E}\left[\frac{1}{X}\right]+\mathbb{E}\left[\rho_X^2(X)\right], \quad (111)$$

where in (110) we have used that

$$\mathbb{E}\left[\left(\frac{Y}{X}-a\right)\Big|X\right]=\frac{\mathbb{E}\left[(Y-aX)\,|X\right]}{X}=0; \quad (112)$$

and in (111) we have used the variance of the Poisson distribution, so that $\mathbb{E}\left[(Y-aX)^2\,|X\right]=aX$. This concludes the proof.

## Appendix E
## Proof of Lemma 6

We first show the upper bound. To that end, let $T$ be a random variable on $[0,1]$ with a pdf given by $f_T(t)=2(1-t)$. Then, $\phi''(u)=\frac{1}{u}$ and

$$\frac{1}{2\Delta_\phi(u,v)}=\frac{1}{\int_0^1\frac{(1-t)}{(1-t)u+tv}\mathrm{d}t} \quad (113)$$

$$=\frac{1}{\frac{1}{2}\mathbb{E}\left[\frac{1}{(1-T)u+Tv}\right]} \quad (114)$$

$$\le 2\mathbb{E}\left[1-T\right]u+2\mathbb{E}\left[T\right]v \quad (115)$$

$$=\frac{4u}{3}+\frac{2v}{3}, \quad (116)$$

where in (115) we have used Jensen's inequality. The proof of the lower bound follows since $v\ge 0$ and

$$\int_0^1\frac{(1-t)}{(1-t)u+tv}\mathrm{d}t\le\frac{1}{u}. \quad (117)$$

This concludes the proof.

## Appendix F
## Proof of Lemma 7

First, observe the following:

$$P_Y(y)=\int\frac{1}{y!}(ax)^y\mathrm{e}^{-ax}\mathrm{d}P_X(x) \quad (118)$$

$$=\frac{a^yP_Y(0)}{y!}\int x^y\frac{\mathrm{e}^{-ax}}{P_Y(0)}\mathrm{d}P_X(x) \quad (119)$$

$$=\frac{a^yP_Y(0)}{y!}\int x^y\mathrm{d}P_Z(x), \quad (120)$$

where in (120) we have defined $\mathrm{d}P_Z(x)=\frac{\mathrm{e}^{-ax}}{P_Y(0)}\mathrm{d}P_X(x)$. The fact that that $P_Z$ is a proper probability distribution follows since

$$\int\mathrm{d}P_Z(x)=\int\frac{\mathrm{e}^{-ax}}{P_Y(0)}\mathrm{d}P_X(x)=\frac{P_Y(0)}{P_Y(0)}=1. \quad (121)$$

Therefore,

$$P_Y(y)=\frac{a^yP_Y(0)}{y!}\mathbb{E}[Z^y] \quad (122)$$

where $Z$ is distributed according to $\mathrm{d}P_Z(x)=\frac{\mathrm{e}^{-ax}}{P_Y(0)}\mathrm{d}P_X(x)$.

Now by using a well-known fact that $\mathbb{E}[X|Y=y]=\frac{1}{a}\frac{(y+1)P_Y(y+1)}{P_Y(y)}$ (see for example [49]), we arrive at

$$\mathbb{E}[X|Y=y]=\frac{1}{a}\frac{(y+1)P_Y(y+1)}{P_Y(y)}=\frac{\mathbb{E}[Z^{y+1}]}{\mathbb{E}[Z^y]}. \quad (123)$$

Therefore, the limits can be bounded as follows:

$$\lim_{y\to\infty}\frac{\mathbb{E}[X|Y=y]}{y+1}=\lim_{y\to\infty}\frac{\mathbb{E}[Z^{y+1}]}{(y+1)\mathbb{E}[Z^y]} \tag{124}$$

$$=\lim_{y\to\infty}\frac{\frac{1}{(y+1)!}\mathbb{E}[Z^{y+1}]}{\frac{1}{y!}\mathbb{E}[Z^y]} \tag{125}$$

$$=\lim_{y\to\infty}\left(\frac{\mathbb{E}[Z^y]}{y!}\right)^{\frac{1}{y}} \tag{126}$$

$$=\lim_{y\to\infty}\left(\frac{\int_0^\infty z^y\frac{e^{-az}}{P_Y(0)}\mathrm{d}P_X(x)}{y!}\right)^{\frac{1}{y}} \tag{127}$$

$$\leq\lim_{y\to\infty}\left(\frac{\frac{1}{a^yP_Y(0)}y^y\mathrm{e}^{-y}}{y!}\right)^{\frac{1}{y}} \tag{128}$$

$$=\frac{1}{a}. \tag{129}$$

where (124) uses the representation in (123); (126) uses the fact that for a positive sequence $\{a_n\}_{n=0}^\infty$ we have that $\lim_{n\to\infty}\frac{a_{n+1}}{a_n}=\lim_{n\to\infty}a_n^{\frac{1}{n}}$; and in (128) follows form $z^ye^{-az}\leq\frac{1}{a^y}y^y\mathrm{e}^{-y}, y\geq 1$.

Furthermore, by using (123) we have that

$$\sup_{y\geq 0}\frac{\mathbb{E}[X|Y=y]}{y+1}=\sup_{n\geq 0}\frac{\mathbb{E}[Z^{y+1}]}{(y+1)\mathbb{E}[Z^y]} \tag{130}$$

Moreover, observe that $\mathbb{E}[Z^y]=\int x^y\frac{\mathrm{e}^{-ax}}{P_Y(0)}\mathrm{d}P_X(x)=\frac{\mathcal{M}_X^{(y)}(-a)}{P_Y(0)}$. Consequently,

$$\sup_{y\geq 0}\frac{\mathbb{E}[X|Y=y]}{y+1}=\sup_{y\geq 0}\frac{\mathcal{M}_X^{(y+1)}(-a)}{(y+1)\mathcal{M}_X^{(y)}(-a)}. \tag{131}$$

This implies that the conditional expectation is bounded by a linear function. Therefore, there must exist numbers $c_1$ and $c_2$ such that

$$\mathbb{E}[X|Y=y]\leq c_1y+c_2,\ y=0,1,\ldots \tag{132}$$

This concludes the proof.

## Appendix G
## Proof of Theorem 6

Using the CR bound in (28) and the expression for the score function in (109) we have that

$$\mathbb{E}\left[\frac{(\frac{\mathrm{d}}{\mathrm{d}X}\log f_{YX}(Y,X))^2}{\Delta_\phi(X,\mathbb{E}[X|Y])}\right]$$

$$=\mathbb{E}\left[\frac{\left(\frac{Y}{X}-a+\rho_X(X)\right)^2}{\Delta_\phi(X,c_1Y+c_2)}\right] \tag{133}$$

$$\leq\mathbb{E}\left[\left(\frac{Y}{X}-a+\rho_X(X)\right)^2\left(\frac{8X}{3}+\frac{4(c_1Y+c_2)}{3}\right)\right], \tag{134}$$

where in the last step we have used the bound in (77). We now compute individual terms in (134).

The first term in (134) is computed as follows:

$$\mathbb{E}\left[\left(\frac{Y}{X}-a+\rho_X(X)\right)^2X\right]$$

$$=\mathbb{E}\left[\left(\frac{Y}{X}-a\right)^2X\right]+2\mathbb{E}\left[\left(\frac{Y}{X}-a\right)\rho(X)X\right]+\mathbb{E}\left[\rho_X^2(X)X\right] \tag{135}$$

$$=\mathbb{E}\left[\left(\frac{Y}{X}-a\right)^2X\right]+\mathbb{E}\left[\rho_X^2(X)X\right] \tag{136}$$

$$=a+\mathbb{E}\left[\rho_X^2(X)X\right], \tag{137}$$

where in (137) we have used that $\mathbb{E}\left[(Y-aX)^2|X\right]=aX$.

The second term in (134) is computed as follows:

$$\mathbb{E}\left[\left(\frac{Y}{X}-a+\rho_X(X)\right)^2Y\right]$$

$$=\mathbb{E}\left[\left(\frac{Y}{X}-a\right)^2Y\right]+2\mathbb{E}\left[\left(\frac{Y}{X}-a\right)\rho_X(X)Y\right]+\mathbb{E}\left[\rho_X^2(X)Y\right] \tag{138}$$

$$=\mathbb{E}\left[\left(\frac{Y}{X}-a\right)^2Y\right]+\mathbb{E}\left[\rho_X^2(X)Y\right] \tag{139}$$

$$=\mathbb{E}\left[\frac{Y^3}{X^2}-2a\frac{Y^2}{X}+a^2Y\right]+a\mathbb{E}\left[\rho_X^2(X)X\right] \tag{140}$$

$$=\mathbb{E}\left[\frac{a^3X^3+3a^2X^2+aX}{X^2}-2a\frac{aX+a^2X^2}{X}+a^3X\right]+a\mathbb{E}\left[\rho_X^2(X)X\right] \tag{141}$$

$$=a^2+a\mathbb{E}\left[\frac{1}{X}\right]+a\mathbb{E}\left[\rho_X^2(X)X\right], \tag{142}$$

where we have used that

$$\mathbb{E}\left[\frac{Y}{X}\rho_X(X)Y\right]=\mathbb{E}\left[\frac{aX+a^2X^2}{X}\rho_X(X)\right]=-a^2, \tag{143}$$

$$\mathbb{E}\left[\rho_X(X)Y\right]=\mathbb{E}\left[\rho_X(X)aX\right]=-a. \tag{144}$$

Furthermore, the third term has been computed in (111) and is given by

$$\mathbb{E}\left[\left(\frac{Y}{X}-a+\rho_X(X)\right)^2\right]=\mathbb{E}\left[\frac{a}{X}\right]+\mathbb{E}\left[\rho_X^2(X)\right]. \tag{145}$$

Finally, combining (134), (137), (142) and (145) leads to the desired bound. This concludes the proof.

## Appendix H
## Proof of Proposition 4

Observe that the MMSE estimator for $X\sim\mathsf{Gam}(\alpha,\theta)$ is given by (72),

$$\frac{\mathbb{V}(X)}{a\mathbb{V}(X)+\mathbb{E}[X]}y+\frac{\mathbb{E}^2[X]}{a\mathbb{V}(X)+\mathbb{E}[X]}=\frac{1}{a+\alpha}y+\frac{\theta}{\theta+\alpha}, \tag{146}$$

where we have used that $\mathbb{E}[X]=\frac{\theta}{\alpha}$ and $\mathbb{V}(X)=\frac{\theta}{\alpha^2}$. Since the two estimators agree, the upper bound is achieved with equality.

The proof of (84) follows by inserting (69) and (70) into (75).

## APPENDIX I
## PROOF OF PROPOSITION 5

The proof of the lower bound follows from Lemma 4 where we have shown that

$$\mathbb{E}[X|Y=y]=\frac{1}{\alpha+a}y+\frac{\theta}{\alpha+a}=c^{\star}y+d^{\star}=c_1y+c_2. \tag{147}$$

From the structure of the BD we have that

$$R_{\phi}(X|Y)=\mathbb{E}\left[X\log X\right]-\mathbb{E}\left[X\log\left(\frac{Y}{\alpha+a}+\frac{\theta}{\alpha+a}\right)\right] -\mathbb{E}\left[X-\frac{Y}{\alpha+a}-\frac{\theta}{\alpha+a}\right]. \tag{148}$$

Now by using Lemma 4 where we have shown that $\mathbb{E}\left[X-\frac{Y}{\alpha+a}-\frac{\theta}{\alpha+a}\right]=0$, and

$$\mathbb{E}\left[X\log X\right]=\frac{\alpha^{\theta}}{\Gamma(\theta)}\int_0^{\infty}x\log(x)\;x^{\theta-1}\mathrm{e}^{-\alpha x}\mathrm{d}x \tag{149}$$

$$=\frac{\theta\left(\log\left(\frac{1}{\alpha}\right)+\psi\left(\theta+1\right)\right)}{\alpha}, \tag{150}$$

where $\psi(t)$ is the digamma function. Next, observe that $x\log x$ is a convex function, and hence,

$$\begin{aligned}
&\mathbb{E}\left[X\log\left(\frac{Y}{\alpha+a}+\frac{\theta}{\alpha+a}\right)\right]\\
&=\mathbb{E}\left[\mathbb{E}[X|Y]\log\left(\frac{Y}{\alpha+a}+\frac{\theta}{\alpha+a}\right)\right]\\
&=\mathbb{E}\left[\left(\frac{Y}{\alpha+a}+\frac{\theta}{\alpha+a}\right)\log\left(\frac{Y}{\alpha+a}+\frac{\theta}{\alpha+a}\right)\right] \qquad (151)\\
&=\mathbb{E}\left[\mathbb{E}\left[\left(\frac{Y}{\alpha+a}+\frac{\theta}{\alpha+a}\right)\log\left(\frac{Y}{\alpha+a}+\frac{\theta}{\alpha+a}\right)|X\right]\right]\\
&\geq\mathbb{E}\left[\left(\frac{\mathbb{E}\left[Y|X\right]}{\alpha+a}+\frac{\theta}{\alpha+a}\right)\log\left(\frac{\mathbb{E}\left[Y|X\right]}{\alpha+a}+\frac{\theta}{\alpha+a}\right)\right] \qquad (152)\\
&=\mathbb{E}\left[\left(\frac{aX}{\alpha+a}+\frac{\theta}{\alpha+a}\right)\log\left(\frac{aX}{\alpha+a}+\frac{\theta}{\alpha+a}\right)\right], \qquad (153)
\end{aligned}$$

where in (152) we have used Jensen's inequality. Observe that (151) leads to (89), the fact that $Y$ is according to a negative-binomial was show in [49].

This concludes the proof.

## ACKNOWLEDGMENT

The authors would like to acknowledge the contribution of Semih Yagli to this work. In particular, the scalar version of Lemma 1 was developed together with him.

## REFERENCES

[1] 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. (SPAWC)*, May 2020, pp. 1–5.

[2] E. Weinstein and A. J. Weiss, "A general class of lower bounds in parameter estimation," *IEEE Trans. Inf. Theory*, vol. IT-34, no. 2, pp. 338–342, Mar. 1988.

[3] H. L. Van Trees, *Detection, Estimation, Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory*. Hoboken, NJ, USA: Wiley, 2004.

[4] J. Ziv and M. Zakai, "Some lower bounds on signal parameter estimation," *IEEE Trans. Inf. Theory*, vol. IT-15, no. 3, pp. 386–391, May 1969.

[5] A. Dytso, M. Faus, A. M. Zoubir, and H. V. Poor, "MMSE bounds for additive noise channels under Kullback–Leibler divergence constraints on the input distribution," *IEEE Trans. Signal Process.*, vol. 67, no. 24, pp. 6352–6367, Dec. 2019.

[6] M. Fauß, A. Dysto, and H. V. Poor, "MMSE bounds under Kullback-Leibler divergence constraints on the joint input-output distribution," 2020, *arXiv:2006.03722*.

[7] K. Watanabe, "Discrete optimal reconstruction distributions for itakura-saito distortion measure," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2399–2404.

[8] D. Guo, S. Shamai (Shitz), 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.

[9] 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.

[10] C. G. Taborda, D. Guo, and F. Perez-Cruz, "Information-estimation relationships over binomial and negative binomial models," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2630–2646, May 2014.

[11] A. Painsky and G. W. Wornell, "Bregman divergence bounds and universality properties of the logarithmic loss," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1658–1673, Mar. 2020.

[12] A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, and J. Lafferty, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, no. 10, pp. 1–45, 2005.

[13] A. Fischer, "Quantization and clustering with Bregman divergences," *J. Multivariate Anal.*, vol. 101, no. 9, pp. 2207–2221, Oct. 2010.

[14] A. Cichocki and S.-I. Amari, "Families of alpha-beta- and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, vol. 12, no. 6, pp. 1532–1568, Jun. 2010.

[15] D. Reem, S. Reich, and A. De Pierro, "Re-examination of Bregman functions and new properties of their divergences," *Optimization*, vol. 68, no. 1, pp. 279–348, Jan. 2019.

[16] M. Broniatowski and W. Stummer, "Some universal insights on divergences for statistics, machine learning and artificial intelligence," in *Geometric Structures of Information*. New York, NY, USA: Springer, 2019, pp. 149–211.

[17] M. Faub, A. Dytso, and H. V. Poor, "An inequality for Bayesian Bregman risks with applications in directional estimation," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst. (MFI)*, Sep. 2021, pp. 1–6.

[18] 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 Comput. Math. Math. Phys.*, vol. 7, no. 3, pp. 200–217, 1967.

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

[20] R. Iyer and J. A. Bilmes, "Submodular-Bregman and the Lovász-Bregman divergences with applications," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2933–2941.

[21] L. Wang, D. E. Carlson, M. R. D. 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.

[22] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Ann. Statist.*, vol. 19, no. 4, pp. 2032–2066, 1991.

[23] 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.

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

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

[26] J. Jiao, K. Venkat, and T. Weissman, "Mutual information, relative entropy and estimation error in semi-martingale channels," *IEEE Trans. Inf. Theory*, vol. 64, no. 10, pp. 6662–6671, Oct. 2018.

[27] G. B. Folland. *Higher-Order Derivatives and Taylor's Formula in Several Variables*. Accessed: Mar. 12, 2019. [Online]. Available: https://sites.math.washington.edu/folland/Math425/taylor2.pdf

[28] L. Li, G. Lebanon, and H. Park, "Fast Bregman divergence NMF using Taylor expansion and coordinate descent," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 307–315.
[29] A. Dytso, M. Fauß, and H. V. Poor, "A Cramér-Rao type bound for Bayesian risk with Bregman loss," 2020, *arXiv:2001.10982*.
[30] E. Lutwak, D. Yang, and G. Zhang, "Cramér–Rao and moment-entropy inequalities for Renyi entropy and generalized Fisher information," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 473–478, Feb. 2005.
[31] J. Naudts, "Estimators, escort probabilities, and $\phi$-exponential families in statistical physics," 2004, *arXiv:math-ph/0402005*.
[32] I. Vajda, "On convergence of information contained in quantized observations," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2163–2172, Aug. 2002.
[33] J.-F. Bercher and C. Vignat, "On minimum Fisher information distributions with restricted support and fixed variance," *Inf. Sci.*, vol. 179, no. 22, pp. 3832–3842, Nov. 2009.
[34] A. Cianchi, E. Lutwak, D. Yang, and G. Zhang, "A unified approach to Cramér–Rao inequalities," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 643–650, Jan. 2014.
[35] M. A. Kumar and K. V. Mishra, "Cramér–Rao lower bounds arising from generalized Csiszár divergences," *Inf. Geometry*, vol. 3, no. 1, pp. 33–59, Jun. 2020.
[36] J.-F. Bercher, "On generalized Cramér–Rao inequalities, generalized Fisher information and characterizations of generalized-Gaussian distributions," *J. Phys. A, Math. Theor.*, vol. 45, no. 25, Jun. 2012, Art. no. 255303.
[37] M. Fauß, A. Dytso, and H. V. Poor, "A variational interpretation of the Cramér–Rao bound," *Signal Process.*, vol. 182, May 2021, Art. no. 107917.
[38] T. Cover and J. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2006.
[39] H. V. Poor, *An Introduction to Signal Detection Estimation*. New York, NY, USA: Springer, 2013.
[40] A. Dytso, R. Bustin, D. Tuninetti, N. Devroye, H. V. Poor, and S. Shamai (Shitz), "On the minimum mean $p$th error in Gaussian noise channels and its applications," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 2012–2037, Dec. 2018.
[41] O. E. Barndorff-Nielsen, *Exponential Families*. Hoboken, NJ, USA: Wileyx, 1980.
[42] P. Diaconis and D. Ylvisaker, "Conjugate priors for exponential families," *Ann. Statist.*, vol. 7, no. 2, pp. 269–281, Mar. 1979.
[43] A. Dytso, M. Faus, and H. V. Poor, "The vector Poisson channel: On the linearity of the conditional mean estimator," *IEEE Trans. Signal Process.*, vol. 68, pp. 5894–5903, 2020.
[44] J. P. Gordon, "Quantum effects in communications systems," *Proc. IRE*, vol. 50, no. 9, pp. 1898–1908, Sep. 1962.
[45] S. Shamai (Shitz), "Capacity of a pulse amplitude modulated direct detection photon channel," *IEE Proc. I. Commun., Speech Vis.*, vol. 137, no. 6, pp. 424–430, Dec. 1990.
[46] C. Lee, C. Lee, and C.-S. Kim, "MMSE nonlocal means denoising algorithm for Poisson noise removal," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2561–2564.
[47] S. Pyatykh and J. Hesser, "MMSE estimation for Poisson noise removal in images," 2015, *arXiv:1512.00717*.
[48] A. Dytso, H. V. Poor, R. Bustin, and S. Shamai (Shitz), "On the structure of the least favorable prior distributions," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1081–1085.
[49] A. Dytso and H. V. Poor, "Estimation in Poisson noise: Properties of the conditional mean estimator," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4304–4323, Jul. 2020.

**Alex Dytso** (Member, IEEE) received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Illinois Chicago, Chicago, in 2016. From September 2016 to August 2020, he was a Post-Doctoral 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). 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 in electrical engineering from Technische Universität München, Germany, in 2010, and the Dr.-Ing. degree in electrical engineering from Technische Universität Darmstadt, Germany, in 2016. In November 2011, he joined the Signal Processing Group, Technische Universität Darmstadt. In September 2019, he joined Prof. H. Vincent Poor's group at Princeton University as a Post-Doctoral Researcher 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. He received the Dissertation Award of the German Information Technology Society for his Ph.D. thesis on robust sequential detection in 2017.

**H. Vincent Poor** (Life Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer science from Princeton University 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 currently the Michael Henry Strater University Professor. From 2006 to 2016, he served as the Dean for Princeton's School of Engineering and Applied Science. He has also held visiting appointments at several other institutions, most recently at Berkeley and Cambridge. His research interests are in the areas of information theory, machine learning and network science, and their applications in wireless networks, energy systems and related fields. Among his publications in these areas is the forthcoming book *Machine Learning and Wireless Communications* (Cambridge University Press). He is a member of the National Academy of Engineering and the National Academy of Sciences, and a foreign member of the Royal Society and other national and international academies. He received the IEEE Alexander Graham Bell Medal in 2017.