A General Derivative Identity for the Conditional Mean Estimator in Gaussian Noise and Some Applications

Alex Dytso*, H. Vincent Poor†, and Shlomo Shamai (Shitz)‡

*,†Department of Electrical Engineering, Princeton University
‡Department of Electrical Engineering, Technion – Israel Institute of Technology
Email: adytso@princeton.edu*, poor@princeton.edu† and sshlomo@ee.technion.ac.il‡

***Abstract*—This paper provides a general derivative identity for the conditional mean estimator of an arbitrary vector signal in Gaussian noise with an arbitrary covariance matrix. This new identity is used to recover and generalize many known identities in the literature and derive some new identities. For example, a new identity is discovered, which shows that an arbitrary higher-order conditional moment is completely determined by the first conditional moment.**

Several applications of the identities are shown. For instance, by using one of the identities, a simple proof of the uniqueness of the conditional mean estimator as a function of the distribution of the signal is shown. Moreover, one of the identities is used to extend the notion of empirical Bayes to higher-order conditional moments. Specifically, based on a random sample of noisy observations, a consistent estimator for a conditional expectation of any order is derived.

***Index Terms*—Conditional mean estimator, empirical Bayes, Gaussian Noise.**

A full version of this paper is accessible at [1].

I. Introduction

There are several derivative identities in the literature that relate conditional mean estimators to other quantities such as the score function and the conditional variance. Such identities are often used in information theory to give way to estimation theoretic arguments (e.g., I-MMSE relationship [2]). In estimation theory such identities are often used to design new estimation procedures (e.g., empirical Bayes [3]). In this work, it is shown that many of the known identities in the literature can be derived systematically from a single unifying derivative identity. Moreover, we use this new identity to derive several generalizations of the previously known identities and discover some new identities. Furthermore, the derived identities are used to propose a generalization of an empirical Bayes procedure.

Contribution: The contribution and the outline of the paper are as follows:

- Section II presents the system model;
- In Section III, Theorem 1 presents a new identity for the Jacobian of the conditional mean;
- Section IV is dedicated to the Tweedie-Robbins-Esposito (TRE) formula. The TRE formula relates the conditional expectation with the score function of the output random variable;
- In Section V, Proposition 1 presents a simple proof of a vector version of Hatsel-Nolte identity, which relates the Jacobian of the conditional expectation to the conditional variance;
- In Section VI, Proposition 2 and Proposition 3 show two vector generalizations of the recursive Jaffer's identity, which relates higher-order conditional moments to the derivatives of the lower-order conditional moments. Moreover, in Section VI-B, Proposition 4 provides an equivalent integral version of Jaffer's identity which shows that every higher-order conditional moment is completely determined by the first-order conditional moment;
- Section VII is dedicated to applications of the derived identities. In Section VII-A, Proposition VII-A shows a small application of Hatsel-Nolte identity concerning the maximum and minimum 'slope' of the conditional mean estimator. In Section VII-B, Theorem 2 uses the new integral generalization of Jaffer's identity to show the uniqueness of the conditional mean estimator as a function of the distribution of the signal. Finally, Section VII-C uses a new integral generalization of Jaffer's identity to extend the notion of empirical Bayes to higher-order conditional moments. Specifically, Theorem 3, based on a random sample of noisy observations, proposes a consistent estimator for a conditional expectation of any order; and
- Section VIII concludes the paper.

Notation: The set of all positive integers is denoted by $\mathbb{N}$, $[n]$ is the set of integers $\{1, \ldots, n\}$, and $\mathbb{R}^n$ is the set of all n-dimensional real-valued vectors.

Deterministic scalar quantities are denoted by lowercase letters, scalar random variables are denoted by uppercase letters, vectors are denoted by bold lowercase letters, random vectors by bold uppercase letters, and matrices by bold uppercase sans serif letters (e.g., x, X, $\mathbf{x}$, $\mathbf{X}$, $\mathbf{X}$). All vectors in the paper are

This work was supported in part by the U. S. National Science Foundation under Grant CCF-1908308 and by the United States-Israel Binational Science Foundation, under Grant BSF-2018710.

column vectors.

The standard basis vectors for $\mathbb{R}^n$ are denoted by $\mathbf{e}_i$, $i \in [n]$. For a matrix A, we use $[\mathsf{A}]_{ij}$ to denote the entry of the row i and column j. The Euclidian norm of a vector $\mathbf{x} \in \mathbb{R}^n$ in this paper is denoted by $\|\mathbf{x}\|$.

The gradient of a function $f : \mathbb{R}^n \to \mathbb{R}$ is denoted by $\nabla_{\mathbf{x}} f(\mathbf{x}) \in \mathbb{R}^n$, and the Jacobian matrix of a function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ is denoted by $\mathsf{J}_{\mathbf{x}}\mathbf{f}(\mathbf{x}) \in \mathbb{R}^{m\times n}$, that is

$$[\mathsf{J}_{\mathbf{x}}\mathbf{f}(\mathbf{x})]_{ij} = \frac{\partial f_i(\mathbf{x})}{\partial x_j}, i \in [m], j \in [n]. \tag{1}$$

The Hadamard product between $\mathsf{A} \in \mathbb{R}^{n\times m}$ and $\mathsf{B} \in \mathbb{R}^{n\times m}$ will be denoted by $\mathsf{A} \odot \mathsf{B}$. Moreover, $\mathsf{A}^{\odot n}$, $n \in \mathbb{N}$ denotes the Hadamard product repeated n times on the matrix A. For a function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, we use the following version of the higher-order gradient $\nabla^{\odot v}\mathbf{f}$, $v \in \mathbb{N}$, which is defined as

$$[\nabla^{\odot v}\mathbf{f}(\mathbf{x})]_i = \frac{\partial^v}{\partial x_i^v}[\mathbf{f}(\mathbf{x})]_i, \, i \in [n]. \tag{2}$$

II. Model

The underlying model considered through this paper is described by the following input-output relationship:

$$\mathbf{Y} = \mathbf{X} + \mathbf{N}, \tag{3}$$

where $\mathbf{N} \in \mathbb{R}^n$ is a zero mean, normally distributed with the covariance matrix $\mathsf{K}_{\mathbf{N}}$, and independent of $\mathbf{X} \in \mathbb{R}^n$. Throughout the paper $\mathsf{K}_{\mathbf{N}}$ is assumed to be a positive definite matrix, and we make no assumptions about the probability distribution of $\mathbf{X}$.

III. A New Identity for the Conditional Expectation

The main result of this paper is the following theorem.

Theorem 1. *Let $\mathbf{U} \in \mathbb{R}^m$ be an arbitrary function of $\mathbf{X} \in \mathbb{R}^n$. Then,*

$$\mathsf{J}_{\mathbf{y}}\mathbb{E}\left[\mathbf{U}|\mathbf{Y}=\mathbf{y}\right] = \mathsf{Cov}(\mathbf{U},\mathbf{X}|\mathbf{Y}=\mathbf{y})\mathsf{K}_{\mathbf{N}}^{-1}, \, \mathbf{y} \in \mathbb{R}^n. \tag{4}$$

Proof: See Appendix. ∎

Conversely, Theorem 1 can be re-written as

$$\mathbb{E}[U_i|\mathbf{Y}=\mathbf{y}] = \mathbb{E}[U_i|\mathbf{Y}=\mathbf{0}] + \oint_{\mathbf{0}}^{\mathbf{y}} \mathsf{Cov}(U_i,\mathbf{X}|\mathbf{Y}=\mathbf{t})\mathsf{K}_{\mathbf{N}}^{-1}\cdot \mathbf{dt}, \tag{5}$$

where $i \in [m]$, and $\oint$ is a line integral over an arbitrary path between $\mathbf{0}$ and $\mathbf{y}$.

We now discuss some consequences of the identity in Theorem 1. Specifically, this would be done by evaluating Theorem 1 with different choices of $\mathbf{U}$ such as $\mathbf{U} = \mathbf{X}$, and $\mathbf{U} = (\mathbf{X}\mathbf{X}^{\mathsf{T}})^{k-1}\mathbf{X}$, $k \in \mathbb{N}$.

IV. Tweedie-Robbins-Esposito Identity

The proof of Theorem 1 will rely on the following identity:

$$\mathbb{E}[\mathbf{X}|\mathbf{Y}=\mathbf{y}] = \mathbf{y} + \mathsf{K}_{\mathbf{N}}\frac{\nabla_{\mathbf{y}} f_{\mathbf{Y}}(\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})}, \tag{6}$$

where $f_{\mathbf{Y}}(\mathbf{y})$ is the probability density function (pdf) of $\mathbf{Y}$. We note that the quantity $\frac{\nabla_{\mathbf{y}} f_{\mathbf{Y}}(\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})}$ is commonly known as the score function.

Literature Review: The scalar version of the identity in (6) has been derived by Robbins in [4] where he credits Maurice Tweedie for the derivation. The vector version of the identity in (6) was derived by Esposito in [5]. Therefore, throughout this paper, we refer to the identity in (6) as Tweedie-Robbins-Esposito identity or TRE for short.

The application of the TRE identity in (6) can considerably simplify the computation of $\mathbb{E}[\mathbf{X}|\mathbf{Y}]$ as we do not need to derive the conditional distribution $P_{\mathbf{X}|\mathbf{Y}}$ and only require to compute $f_{\mathbf{Y}}(\mathbf{y})$ and the gradient of $f_{\mathbf{Y}}(\mathbf{y})$. For an example of such an application, the interested reader is referred to [6] where the TRE identity was used to compute $\mathbb{E}[\mathbf{X}|\mathbf{Y}]$ for the case where $\mathbf{X}$ is uniform on a sphere.

The observation that, via the TRE identity, the conditional expectation can be represented only in terms of the marginal distribution of the output $\mathbf{Y}$ has led to the development of the empirical Bayes procedure [4]; the interested reader is referred to [3] for an overview of the empirical Bayes procedure.

The TRE identity has also been used in the proofs of the scalar and vector versions of the I-MMSE relationship in [7] and [8], respectively.

V. The Identity of Hatsell and Nolte

By setting $\mathbf{U} = \mathbf{X}$ in (4) we arrive at the following identity.

Proposition 1. *For $\mathbf{y} \in \mathbb{R}^n$.*

$$\mathsf{J}_{\mathbf{y}}\mathbb{E}\left[\mathbf{X}|\mathbf{Y}=\mathbf{y}\right] = \mathsf{Var}(\mathbf{X}|\mathbf{Y}=\mathbf{y})\mathsf{K}_{\mathbf{N}}^{-1}. \tag{7}$$

Literature Review: The identity in (7) has been first derived by Hatsell and Nolte in [9] for the case of $\mathsf{K}_{\mathbf{N}} = \mathsf{I}$. The general version in (7) was first derived in [8] where it was used, together with the TRE identity in (6), to give a proof of the vector version of the I-MMSE relationship.

In [10], the scalar version of the identity in (7), was used to show that the minimum mean squared error is Lipschitz continuous with respect to the Wasserstein distance.

VI. Jaffer's Identity

In [11], Jaffer has shown the following identity: for $k \in \mathbb{N} \cup \{0\}$

$$\mathbb{E}\left[X^{k+1}|Y=y\right] = \sigma^2\frac{\mathrm{d}}{\mathrm{d}y}\mathbb{E}\left[X^k|Y=y\right] + \mathbb{E}\left[X^k|Y=y\right]\mathbb{E}[X|Y=y], \tag{8}$$

where σ^2 is the variance of the noise. To the best of our knowledge, Jaffer's identity in (8) has had little applications and is not well known. In what follows, we develop several vector generalizations of Jaffer's identity. Moreover, we derive an alternative but equivalent integral version of the Jaffer's identity and show how this new identity can be used to prove uniqueness of the conditional mean estimator. We also use this integral identity to extend the notion of empirical Bayes to higher order conditional moments.

A. Vector Generalization of Jaffer's Identity

Given the fact that there is no unique generalization of higher moments to the vector case, several vector generalizations of the identity in (8) are possible. Next, we present two such generalizations.

The first generalization of (8) allows to have different exponents across elements of $\mathbf{X}$.

Proposition 2. *For every $m \in [n]$, $v_i \in \mathbb{N} \cup \{0\}$, $i \in [n]$ and $\mathbf{y} \in \mathbb{R}^n$*

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}y_m}\mathbb{E}\left[\prod_{i=1}^{n}\left(\mathbf{e}_i^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}\right)^{v_i}|\mathbf{Y}=\mathbf{y}\right]\\&=\mathbb{E}\left[\prod_{i=1:i\neq m}^{n}\left(\mathbf{e}_i^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}\right)^{v_i}\left(\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}\right)^{v_m+1}|\mathbf{Y}=\mathbf{y}\right]\\&-\mathbb{E}\left[\prod_{i=1}^{n}\left(\mathbf{e}_i^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}\right)^{v_i}|\mathbf{Y}=\mathbf{y}\right]\mathbb{E}\left[\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}|\mathbf{Y}=\mathbf{y}\right].\end{aligned}\tag{9}$$

Proof: The proof follows by evaluating (4) with $\mathbf{U}=\prod_{i=1}^{n}\left(\mathbf{e}_i^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}\right)^{v_i}$. ∎

In the case when $\mathbf{K}_{\mathbf{N}}$ is a diagonal matrix with $[\mathbf{K}_{\mathbf{N}}]_{ii}=\sigma_{ii}^2$ the identity (9) reduces to

$$\begin{aligned}&\mathbb{E}\left[X_m^{v_m+1}\prod_{i=1:i\neq m}^{n}X_i^{v_i}|\mathbf{Y}=\mathbf{y}\right]=\sigma_{ii}^2\frac{\mathrm{d}}{\mathrm{d}y_i}\mathbb{E}\left[\prod_{i=1}^{n}X_i^{v_i}|\mathbf{Y}=\mathbf{y}\right]\\&+\mathbb{E}\left[\prod_{i=1}^{n}X_i^{v_i}|\mathbf{Y}=\mathbf{y}\right]\mathbb{E}[X_m|\mathbf{Y}=\mathbf{y}].\end{aligned}\tag{10}$$

Furthermore, by setting $v_i=0$ for $i\neq m$ the identity in (9) can be re-written in terms of the Hadamard product as follows: for $\mathbf{y}\in\mathbb{R}^n$, $m\in[n]$ and $v\in\mathbb{N}\cup\{0\}$

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}y_m}\mathbb{E}\left[\left(\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}\right)^{\odot v}|\mathbf{Y}=\mathbf{y}\right]=\mathbb{E}\left[\left(\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}\right)^{\odot(v+1)}|\mathbf{Y}=\mathbf{y}\right]\\&-\mathbb{E}\left[\left(\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}\right)^{\odot v}|\mathbf{Y}=\mathbf{y}\right]\odot\mathbb{E}\left[\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}|\mathbf{Y}=\mathbf{y}\right].\end{aligned}\tag{11}$$

The second generalization of (8) is in terms of powers of a matrix.

Proposition 3. *For $k\in\mathbb{N}$ and $\mathbf{y}\in\mathbb{R}^n$*

$$\begin{aligned}&\mathbb{E}\left[(\mathbf{X}\mathbf{X}^{\mathsf{T}})^k|\mathbf{Y}=\mathbf{y}\right]=\mathbf{K}_{\mathbf{N}}\mathbf{J}_{\mathbf{y}}\mathbb{E}\left[(\mathbf{X}\mathbf{X}^{\mathsf{T}})^{k-1}\mathbf{X}|\mathbf{Y}=\mathbf{y}\right]\\&+\mathbb{E}\left[(\mathbf{X}\mathbf{X}^{\mathsf{T}})^{k-1}\mathbf{X}|\mathbf{Y}=\mathbf{y}\right]\mathbb{E}\left[\mathbf{X}^{\mathsf{T}}|\mathbf{Y}=\mathbf{y}\right].\end{aligned}\tag{12}$$

Proof: The proof follows by evaluating (4) with $\mathbf{U}=(\mathbf{X}\mathbf{X}^{\mathsf{T}})^{k-1}\mathbf{X}$. ∎

B. A New Perspective on Jaffer's Identity

Next, we show that Jaffer's identity has an alternative integral version. This new integral representation leads to several interesting consequences. The following auxiliary lemma, the proof of which can be found in [1], would be useful.

Lemma 1. *Let $f_k:\mathbb{R}\to\mathbb{R}$ be a sequence of functions $k=0,1,2,\ldots$ with $f_0\equiv 1$. Then, the recurrence relationship*

$$f_k(x)=\frac{\mathrm{d}}{\mathrm{d}x}f_{k-1}(x)+f_{k-1}(x)f_1(x), k=1,2,\ldots\tag{13}$$

is equivalent to

$$f_k(x)=\mathrm{e}^{-\int_0^x f_1(t)\mathrm{d}t}\frac{\mathrm{d}^k}{\mathrm{d}x^k}\mathrm{e}^{\int_0^x f_1(t)\mathrm{d}t}.\tag{14}$$

Next, we present the integral alternative of Jaffer's identity.

Proposition 4. *Let $\mathbf{f}:\mathbb{R}^n\to\mathbb{R}^n$ and $\mathbf{f}^{-\odot}:\mathbb{R}^n\to\mathbb{R}^n$ be defined as follows: for $i\in[n]$ and $\mathbf{y}\in\mathbb{R}^n$*

$$[\mathbf{f}(\mathbf{y})]_i=\mathrm{e}^{\int_0^{y_i}\mathbb{E}\left[\mathbf{e}_i^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}|\mathbf{Y}=(y_1,\ldots y_{i-1},t,y_{i+1}\ldots,y_n)\right]\mathrm{d}t},\tag{15}$$

$$[\mathbf{f}^{-\odot}(\mathbf{y})]_i=[\mathbf{f}(\mathbf{y})]_i^{-1}.\tag{16}$$

Then, for $k\in\mathbb{N}$ and $\mathbf{y}\in\mathbb{R}^n$

$$\mathbb{E}\left[\left(\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}\right)^{\odot k}|\mathbf{Y}=\mathbf{y}\right]=\mathbf{f}^{-\odot}(\mathbf{y})\odot\nabla^{\odot k}\mathbf{f}(\mathbf{y}).\tag{17}$$

Proof: By setting $v_i=k$ for $i=m$ and $v_i=0$ for $i\neq m$ the identity in (9) reduces to

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}y_m}\mathbb{E}\left[\left(\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}\right)^k|\mathbf{Y}=\mathbf{y}\right]\\&=\mathbb{E}\left[\left(\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}\right)^{k+1}|\mathbf{Y}=\mathbf{y}\right]\\&-\mathbb{E}\left[\left(\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}\right)^k|\mathbf{Y}=\mathbf{y}\right]\mathbb{E}\left[\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}|\mathbf{Y}=\mathbf{y}\right].\end{aligned}\tag{18}$$

Now, by defining

$$f_k(y_m)=\mathbb{E}\left[\left(\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}\right)^k|\mathbf{Y}=\mathbf{y}\right],\tag{19}$$

and applying Lemma 1 to (18) we arrive at

$$\begin{aligned}&\mathbb{E}\left[\left(\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}\right)^k|\mathbf{Y}=\mathbf{y}\right]\\&=\mathrm{e}^{-\int_0^{y_m}\mathbb{E}\left[\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}|\mathbf{Y}=\mathbf{t}\right]\mathrm{d}t_m}\frac{\mathrm{d}^k}{\mathrm{d}y_m^k}\mathrm{e}^{\int_0^{y_m}\mathbb{E}\left[\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}|\mathbf{Y}=\mathbf{t}\right]\mathrm{d}t_m}\end{aligned}\tag{20}$$

where $[\mathbf{t}]_i=y_i$ for $i\neq m$. The proof is concluded by using the definition of the Hadamard product and $\nabla^{\odot v}$ in (2). ∎

In the case when $\mathbf{K}_{\mathbf{N}}$ is a diagonal matrix with $[\mathbf{K}_{\mathbf{N}}]_{ii}=\sigma_i^2$ the identity (17) reduces to: for $y\in\mathbb{R}$ and $k\in\mathbb{N}$

$$\begin{aligned}&\mathbb{E}\left[X_m^k|Y_m=y_m\right]\\&=\sigma_i^{2k}\mathrm{e}^{-\frac{1}{\sigma_m^2}\int_0^{y_m}\mathbb{E}[X_m|Y_m=t]\mathrm{d}t}\frac{\mathrm{d}^k}{\mathrm{d}y_m^k}\mathrm{e}^{\frac{1}{\sigma_m^2}\int_0^{y_m}\mathbb{E}[X_m|Y_m=t]\mathrm{d}t}.\end{aligned}\tag{21}$$

An important feature of the integral version of Jaffer's identity is that any higher-order conditional moment is determined by the first conditional moment. This observation would be used in Section VII-B to show that the conditional expectation is uniquely determined by the distribution of the input X.

Another identity equivalent to that in (21) is shown in the following corollary.

Corollary 1. *For any $k\in\mathbb{N}$ and $y\in\mathbb{R}$*

$$\mathbb{E}[X^k|Y=y]=\sigma^{2k}\frac{\frac{\mathrm{d}^k}{\mathrm{d}y^k}\left(f_Y(y)\mathrm{e}^{\frac{y^2}{2\sigma^2}}\right)}{f_Y(y)\mathrm{e}^{\frac{y^2}{2\sigma^2}}}.\tag{22}$$

Alternatively, let $t \mapsto H_{e_m}(t)$, $m \in [0,1,\ldots]$ be a probabilistic Hermite polynomial, then

$$\mathbb{E}[X^k|Y=y] = \sigma^{2k} \frac{\sum_{m=0}^{k} \binom{k}{m} f_Y^{(k-m)}(y) \frac{(-i)^m}{\sigma^m} H_{e_m}\left(i\frac{y}{\sigma}\right)}{f_Y(y)}. \tag{23}$$

Proof: First, observe that using the scalar TRE identity in (6) we have that

$$\int_0^y \frac{\mathbb{E}\left[X|Y=t\right]}{\sigma^2}\mathrm{d}t = \int_0^y \left(\frac{t}{\sigma^2} + \frac{\mathrm{d}}{\mathrm{d}t}\log(f_Y(t))\right)\mathrm{d}t \tag{24}$$

$$= \frac{y^2}{2\sigma^2} + \log(f_Y(y)) - \log(f_Y(0)). \tag{25}$$

Inserting (25) into (21) concludes the proof of (22). The proof of (23) follows from the generalized product rule. ∎

The identity in (22) can be thought of as a generalization of the TRE identity in (6) to the higher-order moments. Similarly, to the TRE identity the important feature of the identity in (22) is that $\mathbb{E}[X^k|Y]$ depends on the joint distribution $P_{X,Y}$ only through the marginal pdf of Y. In Section VII-C, the identity in (22) will be used to extended the empirical Bayes procedure to higher-order conditional moments.

VII. Applications

In this section, we demonstrate three applications of the above identities. For ease of exposition, we mostly focus on the scalar case.

A. A Small Application of the Identity in (7)

The identity of Hatsell and Nolte in (7) can be used to make various statements about the minimum and maximum 'slope' of $\mathbb{E}\left[\mathbf{X}|\mathbf{Y}\right]$.

Corollary 2. *For every $\mathbf{y} \in \mathbb{R}^n$*

$$0 \leq \mathsf{Tr}\left(\mathbf{J}_{\mathbf{y}}\mathbb{E}\left[\mathbf{X}|\mathbf{Y}=\mathbf{y}\right]\right). \tag{26}$$

In addition, if $\|\mathbf{X}\| \leq R$, then

$$\mathsf{Tr}\left(\mathbf{J}_{\mathbf{y}}\mathbb{E}\left[\mathbf{X}|\mathbf{Y}=\mathbf{y}\right]\right) \leq R^2 \mathsf{Tr}\left(\mathbf{K}_{\mathbf{N}}^{-1}\right). \tag{27}$$

Proof: The proof of the lower bound follows by using (7) together with the properties that both variances are positive definite matrices and that the trace of product of two positive definite matrices is positive. To show the upper bound in (27) we use (7) together with the Cauchy-Schwarz inequality

$$\mathsf{Tr}\left(\mathbf{J}_{\mathbf{y}}\mathbb{E}\left[\mathbf{X}|\mathbf{Y}=\mathbf{y}\right]\right) = \mathsf{Tr}\left(\mathbf{Var}(\mathbf{X}|\mathbf{Y}=\mathbf{y})\mathbf{K}_{\mathbf{N}}^{-1}\right) \tag{28}$$

$$\leq R^2 \mathsf{Tr}\left(\mathbf{K}_{\mathbf{N}}^{-1}\right). \tag{29}$$

∎

B. Uniqueness of the Conditional Expectation via Integral Version of Jaffer's Identity

The identity in (21) leads to an interesting observation that all of the higher conditional moments are completely determined by the first conditional moment. In the next result, we use this observation to establish the uniqueness of the conditional expectation as a function of the distribution of X.

Theorem 2. *The conditional expectation $\mathbb{E}[X|Y]$ is a bijective operator of the distribution of X. In other words, let P_X be the distribution of X, then*

$$\mathbb{E}[X_1|Y_1=y] = \mathbb{E}[X_2|Y_2=y], \forall y \in \mathbb{R} \Longleftrightarrow P_{X_1} = P_{X_2}. \tag{30}$$

Proof: Let $P_{X_1} = P_{X_2}$, then it is immediate that

$$\mathbb{E}[X_1|Y_1=y] = \mathbb{E}[X_2|Y_2=y], \forall y \in \mathbb{R}. \tag{31}$$

Now suppose that $\mathbb{E}[X_1|Y_1=y] = \mathbb{E}[X_2|Y_2=y], \forall y \in \mathbb{R}$, and the goal is to show that $P_{X_1} = P_{X_2}$. First, observe that by the identity in (21) we have that for all $k \in \mathbb{N}$ and all $y \in \mathbb{R}$

$$\mathbb{E}[X_1^k|Y_1=y] = \mathbb{E}[X_2^k|Y_2=y]. \tag{32}$$

We now use (32) to establish that $P_{X_1|Y_1=y} = P_{X_2|Y_2=y}$ for all $y \in \mathbb{R}$. First, without loss of generality assume that $\sigma^2 = 1$. Second, fix some $y \in \mathbb{R}$ in (32) and let

$$m_{1,k} = \mathbb{E}[X_1^k|Y_1=y],\ m_{2,k} = \mathbb{E}[X_2^k|Y_1=y]. \tag{33}$$

The identity (32) implies that $m_{1,k} = m_{2,k}, \forall k \in \mathbb{N}$. The question of whether a distribution of a real valued random variables is determined by its moment is known as Hamburger moment problem [12]. We now use Carleman's sufficient condition to check whether the moments uniquely determine the distribution, this requires verifying that the following sum of moments is divergent:

$$\sum_{k=1}^{\infty} m_{i,2k}^{-\frac{1}{2k}} = \infty. \tag{34}$$

To show that the left side of (34) diverges we will need the following upper bound on the conditional moments shown in [13]: for $i \in \{1,2\}$

$$m_{i,2k} \leq c_i k 2^{k+1}\sqrt{(2k-1)!}, k \in [n] \tag{35}$$

with $c_i = \frac{\mathrm{e}^{\frac{y^2}{2}}}{f_{Y_i}(y)}$. Applying (35) to (34) we have that

$$\sum_{k=1}^{\infty} m_{i,2k}^{-\frac{1}{2k}} \geq \sum_{k=1}^{\infty} \frac{1}{c_i^{\frac{1}{2k}} k^{\frac{1}{2k}} 2^{\frac{k+1}{2k}} ((2k-1)!)^{\frac{1}{4k}}} \tag{36}$$

$$\geq \frac{1}{2c_i^{\frac{1}{2}}} \sum_{k=1}^{\infty} \frac{1}{k^{\frac{1}{2k}} ((2k)!)^{\frac{1}{4k}}}, \tag{37}$$

which diverges by the comparison test. Therefore, the Carleman condition in (34) is satisfied, and the moments determine the distribution. In other words, we have demonstrated that (32) implies that for all $y \in \mathbb{R}$

$$P_{X_1|Y_1=y} = P_{X_2|Y_2=y}. \tag{38}$$

Now the equality in (38) implies that $P_{X_1} = P_{X_2}$. To see this choose some measurable set $A \subseteq \mathbb{R}$ and observe that

$$P_{X_1}(A) = \mathbb{E}[1_A(X_1)] \tag{39}$$

$$= \mathbb{E}\left[\mathbb{E}\left[1_A(X_1)|Y_1\right]\right] \tag{40}$$

$$= \mathbb{E}\left[\mathbb{E}\left[1_A(X_2)|Y_2\right]\right] = P_{X_2}(A). \tag{41}$$

This concludes the proof. ∎

Theorem 2 has been previously shown in [10]. However, our proof here is different than that in [10], and our method is more akin to the proof of the uniqueness of the conditional expectation for the Poisson noise used in [14].

C. Empirical Bayes for Higher-Order Conditional Moments

An interesting application of the original TRE identity is the idea of empirical Bayes proposed by Robbins in [4]. Consider an independent and identically distributed sequence $Y_1, \ldots, Y_n$ according to f_Y, and assume that we have a perfect knowledge of σ. Because the conditional estimator in the TRE formula depends only on the marginal distribution of the output Y, from the Y_i observations, we can build an empirical estimate of $\mathbb{E}[X|Y=y]$ by mimicking the TRE formula

$$\widehat{m}(y) = y + \sigma^2 \frac{\widehat{f}'_Y(y)}{\widehat{f}_Y(y)},\ y \in \mathbb{R}, \tag{42}$$

where $\widehat{f}_Y(y)$ and $\widehat{f}'_Y(y)$ are estimates of $f_Y(y)$ and $f'_Y(y)$, respectively. In other words, we are able to estimate the $\mathbb{E}[X|Y]$ without the knowledge of the prior distribution on X. The interested reader is referred to [3, Chapter 6.1] for a historical account and impact of the empirical Bayes formula.

Let $\widehat{f}_Y^{(k)}(y)$ denote the estimate of $f_Y^{(k)}(y)$ based on a random sample $Y_1, \ldots, Y_n$. Now, inspired by (22), define the estimator of $\mathbb{E}[X^k|Y=y]$ as follows:

$$\widehat{m}_k(y) = \sigma^{2k} \frac{\sum_{m=0}^{k} \binom{k}{m} \widehat{f}_Y^{(k-m)}(y) \frac{(-i)^m}{\sigma^m} H_{e_m}\left(i\frac{y}{\sigma}\right)}{\widehat{f}_Y(y)},\ y \in \mathbb{R}. \tag{43}$$

To estimate $f_Y^{(k)}(y)$ we use the following steps. First, estimate f_Y by using a kernel-density estimator

$$\widehat{f}_Y(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{a} k\left(\frac{t - Y_i}{a}\right), \tag{44}$$

where $a > 0$ is the bandwidth parameter. We take the kernel to be $k(t) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}$. Second, estimate $f_Y^{(k)}(t)$ by taking the derivative of (44) k times.

We conclude this section by showing that the estimator in (43) is consistent.

Theorem 3. *Let $a = \frac{1}{n^u}$ and $t_n = \frac{\sigma^2\sqrt{w\log(n)}}{3}$ for some $u \in \left(0, \frac{1}{2k+6}\right)$ and $w \in (0,u)$. Moreover, assume that $\mathbb{E}[X^2] < \infty$. Then, for every $k \in \mathbb{N}$ and $\sigma^2 > 0$*

$$\lim_{n\to\infty} \mathbb{P}\left[\sup_{|y|\le t_n} \left|\widehat{m}_k(y) - \mathbb{E}[X^k|Y=y]\right| \ge \frac{C_{k,\sigma}}{n^{u-w}}\right] = 0,$$

where $C_{k,\sigma}$ is a constant that depends only on k and σ.

The proof is omitted and can be found in [1].

VIII. Conclusion and Outlook

This work has derived a general derivative identity for a conditional mean estimator. This identity has been used to recover several known derivative identities. Moreover, some generalizations of known identieis and some new identities have been derived. For example, a new integral version of Jaffer's identity has been shown and used to prove the uniqueness of the conditional expectation. Moreover, the identities are used to extend the notion of empirical Bayes to higher-oder moments.

An interesting future direction would be to see if the main identity in Theorem 1 can shed light on the vector generalization of the single crossing property in [15]. It would also be interesting to see if a vector version of integral Jaffer's identity in Proposition 2 can be used to show the uniqueness of the vector conditional expectation as has been done for the scalar case in Theorem 2.

Appendix

First, denote the pdf of $\mathbf{N}$ by $\phi_{\mathbf{K_N}}(\mathbf{n})$ and observe that

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}y_m}\phi_{\mathbf{K_N}}(\mathbf{y}-\mathbf{X}) \\
&= -\frac{1}{2}\phi_{\mathbf{K_N}}(\mathbf{y}-\mathbf{X})\frac{\mathrm{d}}{\mathrm{d}y_m}(\mathbf{y}-\mathbf{X})^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}(\mathbf{y}-\mathbf{X}) && (45)\\
&= \phi_{\mathbf{K_N}}(\mathbf{y}-\mathbf{X})\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}(\mathbf{X}-\mathbf{y}). && (46)
\end{aligned}$$

Second, observe the following sequence of steps:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}y_m}\mathbb{E}\left[\mathbf{U}|\mathbf{Y}=\mathbf{y}\right] \\
&= \frac{\mathrm{d}}{\mathrm{d}y_m}\mathbb{E}\left[\mathbf{U}\frac{\phi_{\mathbf{K_N}}(\mathbf{y}-\mathbf{X})}{f_{\mathbf{Y}}(\mathbf{y})}\right] && (47)\\
&= \mathbb{E}\left[\mathbf{U}\frac{\mathrm{d}}{\mathrm{d}y_m}\frac{\phi_{\mathbf{K_N}}(\mathbf{y}-\mathbf{X})}{f_{\mathbf{Y}}(\mathbf{y})}\right] && (48)\\
&= \mathbb{E}\left[\mathbf{U}\frac{\frac{\mathrm{d}}{\mathrm{d}y_m}\phi_{\mathbf{K_N}}(\mathbf{y}-\mathbf{X})}{f_{\mathbf{Y}}(\mathbf{y})}\right] - \mathbb{E}\left[\mathbf{U}|\mathbf{Y}=\mathbf{y}\right]\frac{\frac{\mathrm{d}}{\mathrm{d}y_m}f_{\mathbf{Y}}(\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})} && (49)\\
&= \mathbb{E}\left[\mathbf{U}\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}|\mathbf{Y}=\mathbf{y}\right] \\
&\quad - \mathbb{E}\left[\mathbf{U}|\mathbf{Y}=\mathbf{y}\right]\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{y} - \mathbb{E}\left[\mathbf{U}|\mathbf{Y}=\mathbf{y}\right]\frac{\frac{\mathrm{d}}{\mathrm{d}y_m}f_{\mathbf{Y}}(\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})} && (50)\\
&= \mathbb{E}\left[\mathbf{U}\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}|\mathbf{Y}=\mathbf{y}\right] \\
&\quad - \mathbb{E}\left[\mathbf{U}|\mathbf{Y}=\mathbf{y}\right]\mathbf{e}_m^{\mathsf{T}}\left(\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{y} + \frac{\nabla_{\mathbf{y}}f_{\mathbf{Y}}(\mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})}\right) && (51)\\
&= \mathbb{E}\left[\mathbf{U}\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}|\mathbf{Y}=\mathbf{y}\right] \\
&\quad - \mathbb{E}\left[\mathbf{U}|\mathbf{Y}=\mathbf{y}\right]\mathbb{E}\left[\mathbf{e}_m^{\mathsf{T}}\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{X}|\mathbf{Y}=\mathbf{y}\right] && (52)\\
&= \mathbb{E}\left[\mathbf{U}\mathbf{X}^{\mathsf{T}}|\mathbf{Y}=\mathbf{y}\right]\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{e}_m \\
&\quad - \mathbb{E}\left[\mathbf{U}|\mathbf{Y}=\mathbf{y}\right]\mathbb{E}\left[\mathbf{X}^{\mathsf{T}}|\mathbf{Y}=\mathbf{y}\right]\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{e}_m && (53)\\
&= \mathbf{Cov}(\mathbf{U},\mathbf{X}|\mathbf{Y}=\mathbf{y})\mathbf{K}_{\mathbf{N}}^{-1}\mathbf{e}_m, && (54)
\end{aligned}$$

where the equalities follow from: (47) using Bayes' theorem; (50) using the expression in (46); and (52) using the TRE identity in (6). This concludes the proof.

REFERENCES

[1] A. Dytso, H. V. Poor, and S. Shamai (Shitz), "A general derivative identity for the conditional mean estimator in Gaussian noise and some applications," 2020. [Online]. Available: http://www.princeton.edu/\%7Eadytso/papers/EB.pdf

[2] D. Guo, S. Shamai, and S. Verdú, *The Interplay Between Information and Estimation Measures*. now Publishers Incorporated, 2013.

[3] B. Efron and T. Hastie, *Computer Age Statistical Inference*. Cambridge University Press, 2016, vol. 5.

[4] H. Robbins, "An empirical Bayes approach to statistics," in *Proc. Third Berkeley Symp. Math Statist. Probab.* Citeseer, 1956.

[5] R. Esposito, "On a relation between detection and estimation in decision theory," *Inf. Control*, vol. 12, no. 2, pp. 116–120, February 1968.

[6] A. Dytso, M. Al, H. V. Poor, and S. Shamai (Shitz), "On the capacity of the peak power constrained vector Gaussian channel: An estimation theoretic perspective," *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3907–3921, 2019.

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

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

[9] C. Hatsell and L. Nolte, "Some geometric properties of the likelihood ratio (corresp.)," *IEEE Trans. Inf. Theory*, vol. 17, no. 5, pp. 616–618, 1971.

[10] Y. Wu and S. Verdú, "Functional properties of minimum mean-square error and mutual information," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1289–1301, 2012.

[11] A. G. Jaffer, "A note on conditional moments of random signals in Gaussian noise (corresp.)," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 513–514, 1972.

[12] J. A. Shohat and J. D. Tamarkin, *The Problem of Moments*. American Mathematical Soc., 1943, no. 1.

[13] D. Guo, Y. Wu, S. Shamai, and S. Verdú, "Estimation in Gaussian noise: Properties of the minimum mean-square error," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2371–2385, 2011.

[14] A. Dytso and H. V. Poor, "Estimation in Poisson noise: Properties of the conditional mean estimator," *IEEE Trans. Inf. Theory*, 2020, to appear.

[15] R. Bustin, M. Payaró, D. P. Palomar, and S. Shamai, "On MMSE crossing properties and implications in parallel vector Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 59, no. 2, pp. 818–844, 2013.