Polynomial Approximations of Conditional Expectations in Scalar Gaussian Channels

Wael Alghamdi and Flavio P. Calmon
Harvard University
alghamdi@g.harvard.edu, flavio@seas.harvard.edu

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9517932

***Abstract*—We consider a channel $Y = X + N$ where X is a random variable satisfying $\mathbb{E}[|X|] < \infty$ and N is an independent standard normal random variable. We show that the minimum mean-square estimator of X from Y, which is given by the conditional expectation $\mathbb{E}[X \mid Y]$, is a polynomial in Y if and only if it is linear or constant; these two cases correspond to X being Gaussian or a constant, respectively. We also prove that the higher-order derivatives of $y \mapsto \mathbb{E}[X \mid Y = y]$ are expressible as multivariate polynomials in the functions $y \mapsto \mathbb{E}\left[(X - \mathbb{E}[X \mid Y])^k \mid Y = y\right]$ for $k \in \mathbb{N}$. These expressions yield bounds on the 2-norm of the derivatives of the conditional expectation. These bounds imply that, if X has a compactly-supported density that is even and decreasing on the positive half-line, then the error in approximating the conditional expectation $\mathbb{E}[X \mid Y]$ by polynomials in Y of degree at most n decays faster than any polynomial in n.**

I. Introduction

We investigate the extent to which polynomials can approximate conditional expectations in the scalar Gaussian channel. For

$$Y = X + N, \tag{1}$$

where X has finite variance and $N \sim \mathcal{N}(0,1)$ is independent of X, the conditional expectation $\mathbb{E}[X \mid Y]$ is the minimum mean-square error (MMSE) estimator:

$$\min_{Z} \ \mathbb{E}\left[|X - Z|^2\right] = \mathbb{E}\left[|X - \mathbb{E}[X \mid Y]|^2\right], \tag{2}$$

where the minimization is taken over all $\sigma(Y)$-measurable random variables Z. It is well-known that $\mathbb{E}[X \mid Y]$ is linear (i.e., a first degree polynomial in Y) if and only if X is Gaussian (see, e.g., [1]). We take this a step further and examine when $\mathbb{E}[X \mid Y]$ is close to being a polynomial. Specifically, we focus on two questions:

(Q1) For which distributions of X is a polynomial estimator optimal (in the mean-square sense) for reconstructing X from Y?

(Q2) When the MMSE estimator $\mathbb{E}[X \mid Y]$ is not a polynomial, how well can it be approximated by a polynomial?

In the course of answering **(Q2)**, we answer another fundamental question:

(Q3) How can the higher-order derivatives of $\mathbb{E}[X \mid Y = y]$ in y be expressed and bounded?

We provide a full answer for **(Q1)** in Theorem 1, where we show that the MMSE estimator is a polynomial if and only if X is Gaussian or constant. In other words, the only way $\mathbb{E}[X \mid Y]$ can be a polynomial is if it is linear in Y or is a constant.

For the second question, if X has a probability density function (PDF) or a probability mass function (PMF) p_X that is compactly-supported, even, and decreasing over $[0, \infty) \cap \operatorname{supp}(p_X)$, then we show in Theorem 3 that for all positive integers n and k satisfying $n \geq \max(k-1, 1)$ we have that

$$\inf_{q \in \mathscr{P}_n} \|\mathbb{E}[X \mid Y] - q(Y)\|_2 = O_{X,k}\left(\frac{1}{n^{k/2}}\right). \tag{3}$$

Here, $\mathscr{P}_n$ denotes the set of all polynomials with real coefficients of degree at most n, the implicit constant in (3) can depend on X and k, and $\|\cdot\|_2$ refers to the P_Y-weighted 2-norm, i.e., $\|f(Y)\|_2^2 = \mathbb{E}[f(Y)^2]$.

The result in (3) hinges on our answer to **(Q3)** in virtue of it giving a uniform upper bound on the derivatives of the conditional expectation (see Theorem 2): there are absolute constants $\{\eta_k\}_{k \geq 1}$ such that

$$\sup_{\mathbb{E}[|X|] < \infty} \left\| \frac{d^k}{dy^k} \, \mathbb{E}[X \mid Y = y] \right\|_2 \leq \eta_k. \tag{4}$$

The bound in (4) is a corollary of our answer to the other half of **(Q3)**, where we express the derivatives of the conditional expectation in the form (see Proposition 1)

$$\frac{d^{r-1}}{dy^{r-1}} \, \mathbb{E}[X \mid Y = y] = \sum_{\substack{2\lambda_2 + \cdots + r\lambda_r = r \\ \lambda_2, \cdots, \lambda_r \in \mathbb{N}}} e_{\lambda_2, \cdots, \lambda_r} \prod_{i=2}^{r} \mathbb{E}\left[(X - \mathbb{E}[X \mid Y])^i \mid Y = y\right]^{\lambda_i} \tag{5}$$

for some explicit integers $e_{\lambda_2, \cdots, \lambda_r}$ that we define in the sequel. Setting $r = 2$ in (5) recovers the first derivative [2]

$$\frac{d}{dy} \, \mathbb{E}[X \mid Y = y] = \operatorname{Var}[X \mid Y = y]. \tag{6}$$

These results complement our previous work in [3], where we show that if X has a moment generating function (MGF), then there are constants $\{c_{n,j}\}_{n \in \mathbb{N}, j \in [n]}$ such that

$$\mathbb{E}[X \mid Y] = \lim_{n \to \infty} \sum_{j \in [n]} c_{n,j} Y^j \tag{7}$$

holds in the mean-square sense. In fact, we may choose

$$(c_{n,0}, \cdots, c_{n,n}) = \mathbb{E}\left[(X, XY, \cdots, XY^n)\right] \boldsymbol{M}_{Y,n}^{-1} \tag{8}$$

where the Hankel matrix of moments of Y is denoted by

$$\boldsymbol{M}_{Y,n} := \left(\mathbb{E}\left[Y^{i+j}\right]\right)_{(i,j)\in[n]^2}. \tag{9}$$

Denoting $\boldsymbol{Y}^{(n)} = (1, Y, \cdots, Y^n)^T$, the polynomial

$$E_n[X \mid Y] = \mathbb{E}\left[(X, XY, \cdots, XY^n)\right] \boldsymbol{M}_{Y,n}^{-1} \boldsymbol{Y}^{(n)} \tag{10}$$

is the orthogonal projection of $\mathbb{E}[X \mid Y]$ onto the subspace $\mathscr{P}_n(Y) := \{p(Y) \mid p \in \mathscr{P}_n\}$. This projection characterization, in turn, makes $E_n[X \mid Y]$ the best-polynomial approximation (in the weighted L^2-norm sense) of the conditional expectation $\mathbb{E}[X \mid Y]$. Specifically, $E_n[X \mid Y]$ uniquely solves the approximation problem

$$E_n[X \mid Y] = \operatorname*{argmin}_{q(Y)\in\mathscr{P}_n(Y)} \|q(Y) - \mathbb{E}[X \mid Y]\|_2. \tag{11}$$

For (3), we apply solutions to the Bernstein approximation problem (see [4] for a comprehensive survey). The original Bernstein approximation problem extends Weierstrass approximation to polynomial approximation in $L^\infty(\mathbb{R}, \mu)$ for a measure μ that is absolutely continuous with respect to the Lebesgue measure. The work by Ditzian and Totik [5]—which introduces moduli of smoothness—shows that tools used to solve the Bernstein approximation problem can also be useful for polynomials approximation in $L^p(\mathbb{R}, \mu)$ for all $p \geq 1$. We apply their results for the case $p = 2$.

MMSE estimation in Gaussian channels plays a central role in several information-theoretic applications (e.g., [1, 6–9]). The MMSE dimension [10] is a measure of nonlinearity of the MMSE estimator. The first-order derivative of the conditional expectation in Gaussian channels has been treated in [2]. In particular, formula (6) is generalized in [2] to the multivariate case.

The bound in (3) induces a bound on the gap between the MSE achieved by polynomial estimators and the MMSE. Indeed, the loss from replacing the MMSE estimator $\mathbb{E}[X \mid Y]$ with its best-polynomial approximation $E_n[X \mid Y]$ is

$$\Delta_{n,X} := \|X - E_n[X \mid Y]\|_2^2 - \|X - \mathbb{E}[X \mid Y]\|_2^2 \tag{12}$$

$$= \|E_n[X \mid Y] - \mathbb{E}[X \mid Y]\|_2^2, \tag{13}$$

where (13) follows by the orthogonality principle for the conditional expectation $\mathbb{E}[X \mid Y]$. Hence, inequality (3) yields the bounds $\Delta_{n,X} = O_{X,\ell}(n^{-\ell})$ for every fixed $\ell > 0$. We note that utilizing higher-order polynomials as proxies of the MMSE has appeared, e.g., in approaches to denoising [11].

The full proofs can be found in the extended version of this paper [12].

A. Notation

The probability measure induced by a random variable (RV) X is denoted by P_X. If X is continuous (resp. discrete), then its PDF (resp. PMF) is denoted by p_X. We use the notation $\|\cdot\|_q$ for norms of RVs, i.e., for $q \geq 1$ we have $\|X\|_q^q = \mathbb{E}\left[|X|^q\right]$. We say that a RV X is n-times integrable if it satisfies $\|X\|_n < \infty$, and it is integrable if $\|X\|_1 < \infty$. The norm of the Banach space $L^q(\mathbb{R})$ (for $q \geq 1$) is denoted by $\|\cdot\|_{L^q(\mathbb{R})}$.

The characteristic function of a RV Z is denoted by $\varphi_Z(t) := \mathbb{E}\left[e^{itZ}\right]$. We let $\mathscr{P}_n$ denote the set of polynomials of degree at most n with real coefficients. For $n \in \mathbb{N}$, we set $[n] := \{0, 1, \cdots, n\}$ and denote the set of all finite-length tuples of non-negative integers by $\mathbb{N}^*$.

For every integer $r \geq 2$, let Π_r be the set of unordered integer partitions $r = r_1 + \cdots + r_k$ of r into integers $r_j \geq 2$. We encode Π_r via a list of the multiplicities of the parts as

$$\Pi_r := \{(\lambda_2, \cdots, \lambda_\ell) \in \mathbb{N}^* \ ; \ 2\lambda_2 + \cdots + \ell\lambda_\ell = r\}. \tag{14}$$

In (14), $\ell \geq 2$ is free, and trailing zeros are ignored (i.e., $\lambda_\ell > 0$). For a partition $(\lambda_2, \cdots, \lambda_\ell) = \boldsymbol{\lambda} \in \Pi_r$ having $m = \lambda_2 + \cdots + \lambda_\ell$ parts, we denote[1]

$$c_{\boldsymbol{\lambda}} := \frac{1}{m}\binom{m}{\lambda_2, \cdots, \lambda_\ell}\binom{r}{\underbrace{2, \cdots, 2}_{\lambda_2}; \cdots; \underbrace{\ell, \cdots, \ell}_{\lambda_\ell}} \tag{15}$$

and

$$e_{\boldsymbol{\lambda}} := (-1)^{m-1} c_{\boldsymbol{\lambda}}. \tag{16}$$

We set[2] $C_r := \sum_{\boldsymbol{\lambda}\in\Pi_r} c_{\boldsymbol{\lambda}}$. Let $\left\{{r \atop m}\right\}$ denote the Stirling numbers of the second kind (i.e., the number of unordered set-partitions of an r-element set into m nonempty subsets). The integer C_r can be expressed as

$$C_r = \sum_{k=1}^{r} (k-1)! \sum_{j=0}^{k} (-1)^j \binom{r}{j} \left\{{r-j \atop k-j}\right\}. \tag{17}$$

A derivation of this formula is included in [12]. The first few values of C_r (for $2 \leq r \leq 7$) are $1, 1, 4, 11, 56, 267$, and we have the asymptotic $C_r \sim (r-1)!/\alpha^r$ for some constant $\alpha \approx 1.146$ as $r \to \infty$ (see [13]) and the crude bound $C_r < r^r$.

B. Assumptions

We assume only that X is integrable and $N \sim \mathcal{N}(0, 1)$ is independent of X to prove that the conditional expectation $\mathbb{E}[X \mid X + N]$ cannot be a polynomial of degree exceeding 1 (Theorem 1) and derive the formula for the higher-order derivatives of the conditional expectation (Proposition 1) along with the ensuing bounds on the norms of the derivatives (Theorem 2). For the Bernstein approximation theorem we prove for $\mathbb{E}[X \mid X+N]$ (Theorem 3), we impose the additional assumption that X is either continuous or discrete with a PDF or a PMF belonging to the set we define next.

Definition 1. Let $\mathscr{D}$ denote the set of compactly-supported even PDFs or PMFs p that are non-increasing over $[0, \infty) \cap \operatorname{supp}(p)$.

[1]The integer $c_{\boldsymbol{\lambda}}$ counts the number of cyclically-invariant ordered set-partitions of an r-element set into $m = \lambda_2 + \cdots + \lambda_\ell$ subsets where, for each $k \in \{2, \cdots, \ell\}$, exactly λ_k parts have size k.

[2]The integer C_r counts the total number of cyclically-invariant ordered set-partitions of an r-element set into subsets of sizes at least 2.

II. Polynomial Conditional Expectation

We start by showing that the only way $\mathbb{E}[X \mid Y]$ can be a polynomial, for integrable X and $Y = X + N$ a Gaussian perturbation, is if X is Gaussian or constant. The proof is carried in two steps. First, we show that a degree-m non-constant polynomial $\mathbb{E}[X \mid Y]$ requires $p_Y = e^{-h}$ for some polynomial h with $\deg h = m+1$. The second step is showing that, because $p_Y = e^{-h}$ is a convolution of the Gaussian kernel, $m = 1$.

The following lemma will be useful, and we include its proof in the extended version of this paper [12].

Lemma 1. *For a RV X and a polynomial p, if $p(X)$ is integrable then so is $X^{\deg(p)}$.*

This lemma will allow us to conclude the finiteness of all moments of X directly from the hypotheses that $\mathbb{E}[X \mid Y]$ is a polynomial of degree exceeding 1 and $\|X\|_1 < \infty$, because $\|\mathbb{E}[X \mid Y]\|_k \le \|X\|_k$ for every $k \ge 1$.

Theorem 1. *For $Y = X+N$ where X is an integrable RV and $N \sim \mathcal{N}(0,1)$ independent of X, the conditional expectation $\mathbb{E}[X \mid Y]$ cannot be a polynomial in Y with degree greater than 1. Therefore, the MMSE estimator in a Gaussian channel with finite-variance input is a polynomial if and only if the input is Gaussian or constant.*

Proof. Suppose, for the sake of contradiction, that

$$\mathbb{E}[X \mid Y] = q(Y) \tag{18}$$

for some polynomial with real coefficients q of degree $m := \deg q > 1$. The contradiction we derive will be that the probability measure defined by

$$Q(B) := \frac{1}{a}\int_B e^{-x^2/2}\, dP_X(x) \tag{19}$$

for every Borel subset $B \subset \mathbb{R}$, where $a = \mathbb{E}[e^{-X^2/2}]$ is the normalization constant, would necessarily have a cumulant generating function that is a polynomial of degree $m+1 > 2$. Let R be a RV distributed according to Q. We note that the polynomial q is uniquely determined by (18) because Y is continuous, for if $q(Y) = g(Y)$ for a polynomial g then the support of Y must be a subset of the roots of $q - g$.

The proof strategy is to compute the PDF p_Y in two ways. One way is to compute p_Y as a convolution

$$p_Y(y) = \frac{1}{\sqrt{2\pi}}\mathbb{E}\left[e^{-(X-y)^2/2}\right]. \tag{20}$$

This equation shows by Lebesgue's dominated convergence that p_Y is continuous. The second way to compute p_Y is via the inverse Fourier transform of φ_Y. We consider the Fourier transform that takes an integrable function φ to $\widehat{\varphi}(y) := \int_{\mathbb{R}} \varphi(t)e^{-iyt}\, dt$, so the inverse Fourier transform takes an integrable function p to $(2\pi)^{-1}\int_{\mathbb{R}} p(y)e^{ity}\, dy$. Now, $\varphi_Y = \varphi_X\varphi_N$ is integrable; indeed, $|\varphi_X| \le 1$ and $\varphi_N(t) = e^{-t^2/2}$. Also, being a characteristic function, φ_Y is continuous too. Therefore, by the Fourier inversion theorem, since $\varphi_Y/(2\pi)$ is the inverse Fourier transform of p_Y, we obtain that $p_Y = \widehat{\varphi_Y}/(2\pi)$. Equating this latter equation with (20), then multiplying both sides by $\sqrt{2\pi}e^{y^2/2}/a$, that $R \sim Q$ (see (19)) implies

$$\mathbb{E}\left[e^{Ry}\right] = \frac{1}{a\sqrt{2\pi}}e^{y^2/2}\widehat{\varphi_Y}(y). \tag{21}$$

Equation (21) holds for every $y \in \mathbb{R}$. The rest of the proof derives a contradiction by showing that $\widehat{\varphi_Y} = e^G$ for some polynomial G of degree $m+1 > 2$.

Integrability of X implies integrability of $\mathbb{E}[X \mid Y]$, so for every $t \in \mathbb{R}$

$$\mathbb{E}\left[e^{itY}(X - \mathbb{E}[X \mid Y])\right] = 0. \tag{22}$$

Substituting $X = Y - N$ and $\mathbb{E}[X \mid Y] = q(Y)$ into (22),

$$\mathbb{E}\left[e^{itY}(Y - N - q(Y))\right] = 0. \tag{23}$$

Because the RVs $e^{itY}(Y - q(Y))$ and $e^{itY}N$ are integrable, we may split the expectation to obtain

$$\mathbb{E}\left[e^{itY}(Y - q(Y))\right] - \mathbb{E}\left[e^{itY}N\right] = 0. \tag{24}$$

We rewrite equation (24) in terms of the characteristic functions of Y and N.

Since $q(Y)$ is integrable, Lemma 1 implies that Y is m-times integrable. In particular, $\mathbb{E}\left[|(X+z)^m|\right] < \infty$ for some $z \in \mathbb{R}$. By Lemma 1 again, X is m-times integrable. Hence, for each $k \in [m]$ and $Z \in \{X, N, Y\}$, that $\mathbb{E}\left[|Z|^k\right] < \infty$ implies that the k-th derivative $\varphi_Z^{(k)}$ exists everywhere and

$$(-i)^k\varphi_Z^{(k)}(t) = \mathbb{E}\left[e^{itZ}Z^k\right]. \tag{25}$$

For the term $\mathbb{E}\left[e^{itY}N\right]$ in (24), plugging in $Y = X + N$, we infer from (25) that

$$\mathbb{E}\left[e^{itY}N\right] = \varphi_X(t)\mathbb{E}\left[e^{itN}N\right] = -i\varphi_X(t)\varphi_N'(t). \tag{26}$$

But $\varphi_N(t) = e^{-t^2/2}$, so $\varphi_N'(t) = -t\varphi_N(t)$, hence (26) yields

$$\mathbb{E}\left[e^{itY}N\right] = it\varphi_X(t)\varphi_N(t) = it\varphi_Y(t). \tag{27}$$

Let α_k for $k \in [m]$ be real constants such that $q(u) = \sum_{k\in[m]}\alpha_k u^k$ identically over $\mathbb{R}$, so $\alpha_m \neq 0$. For the first term in (24), utilizing (25) repeatedly we obtain

$$\mathbb{E}\left[e^{itY}(Y - q(Y))\right] = -i\sum_{k\in[m]} c_k\varphi_Y^{(k)}(t) \tag{28}$$

where we define the constants

$$c_k := (-i)^{k+1}\alpha_k + \delta_{1,k} = \begin{cases} (-i)^{k+1}\alpha_k & \text{if } k \in [m]\setminus\{1\}, \\ 1 - \alpha_1 & \text{if } k = 1. \end{cases} \tag{29}$$

Plugging (27) and (28) in (24), we get the differential equation

$$t\varphi_Y(t) + \sum_{k\in[m]} c_k\varphi_Y^{(k)}(t) = 0. \tag{30}$$

We will transform the differential equation (30) into a linear differential equation in the Fourier transform of φ_Y. For this, we need first to show that for each $k \in [m]$ the derivative $\varphi_Y^{(k)}$ is integrable so that its Fourier transform is well-defined.

Now, repeated differentiation of $\varphi_Y(t) = \varphi_X(t)e^{-t^2/2}$ shows that for each $k \in [m]$ there is a polynomial r_k in $k+2$ variables such that

$$\varphi_Y^{(k)}(t) = r_k\left(t, \varphi_X(t), \varphi_X'(t), \cdots, \varphi_X^{(k)}(t)\right)e^{-t^2/2}. \tag{31}$$

Indeed, we start with $r_0(t,u) = u$ because $\varphi_Y(t) = \varphi_X(t)e^{-t^2/2}$. Now, suppose (31) holds for some $k \in [m-1]$. The derivative (with respect to t) of the r_k term is

$$\frac{d}{dt} r_k\left(t, \varphi_X(t), \cdots, \varphi_X^{(k)}(t)\right) = s_k\left(t, \varphi_X(t), \cdots, \varphi_X^{(k+1)}(t)\right) \tag{32}$$

for some polynomial s_k in $k+3$ variables. Therefore, differentiating (31), we get

$$\varphi_Y^{(k+1)}(t) = r_{k+1}\left(t, \varphi_X(t), \varphi_X'(t), \cdots, \varphi_X^{(k+1)}(t)\right) e^{-t^2/2} \tag{33}$$

where

$$\begin{aligned} r_{k+1}\left(t, u_0, \cdots, u_{k+1}\right) &:= s_k\left(t, u_0, \cdots, u_{k+1}\right) \\ &\quad - t \cdot r_k\left(t, u_0, \cdots, u_k\right) \end{aligned} \tag{34}$$

is a polynomial in $k+3$ variables. Therefore (31) holds for all $k \in [m]$. Now, for each $j \in [m]$, we have by (25) the uniform bound $|\varphi_X^{(j)}(t)| \le \mathbb{E}\left[|X|^j\right]$. Therefore, for each $k \in [m]$, letting v_k be the same polynomial as r_k but with the coefficients replaced with their absolute values, the triangle inequality applied to (31) yields the bound $|\varphi_Y^{(k)}(t)| \le \eta_k(t)e^{-t^2/2}$ where $\eta_k(t) := v_k\left(|t|, 1, \mathbb{E}[|X|], \cdots, \mathbb{E}\left[|X|^k\right]\right)$ is a (positive) polynomial in $|t|$. Since $\int_{\mathbb{R}} \eta_k(t)e^{-t^2/2}\,dt < \infty$, we obtain that $\varphi_Y^{(k)}$ is integrable for each $k \in [m]$.

Taking the Fourier transform in the differential equation (30) we infer

$$i\widehat{\varphi_Y}'(y) + \widehat{\varphi_Y}(y) \sum_{k\in[m]} c_k (iy)^k = 0. \tag{35}$$

We rewrite this equation in terms of the α_k (see (29)) as

$$\widehat{\varphi_Y}'(y) - \widehat{\varphi_Y}(y) \sum_{k\in[m]} (\alpha_k - \delta_{1,k}) y^k = 0. \tag{36}$$

Equation (35) necessarily implies

$$\widehat{\varphi_Y}(y) = D \exp\left(\sum_{k\in[m]} \frac{\alpha_k - \delta_{1,k}}{k+1} y^{k+1}\right) \tag{37}$$

for some constant D. Since $p_Y = \widehat{\varphi_Y}/(2\pi)$, we necessarily have $D > 0$. Therefore, we obtain the desired form for $\widehat{\varphi_Y}$, namely, $\widehat{\varphi_Y} = e^G$ where $G \in \mathscr{P}_{m+1} \setminus \mathscr{P}_m$ is given by[3]

$$G(y) := \sum_{k\in[m]} \frac{\alpha_k - \delta_{1,k}}{k+1} y^{k+1} + \log(D). \tag{38}$$

Plugging in this formula for $\widehat{\varphi_Y}$ in (21), we obtain that the cumulant-generating function of the RV R is the degree-$(m+1)$ polynomial $G(y) + y^2/2 - \log(a\sqrt{2\pi})$, contradicting Marcinkiewicz's theorem that a cumulant-generating function has degree at most 2 if it were a polynomial (see, e.g., [14, Theorem 2.5.3]). This concludes the proof by contradiction that $\mathbb{E}[X \mid Y]$ cannot be a polynomial of degree at least 2.

For the second statement in the theorem, we consider the remaining two cases that $\mathbb{E}[X \mid Y]$ is a linear expression in Y or is a constant. If $\mathbb{E}[X \mid Y]$ is constant, then differentiating and taking the expectation in (6) yields that $\|X - \mathbb{E}[X \mid Y]\|_2 = 0$, i.e., $X = \mathbb{E}[X \mid Y]$ is constant. Finally, under the assumption that X has finite variance, $\mathbb{E}[X \mid Y]$ is linear if and only if X is Gaussian (see, e.g., [1]). We note that if one requires only that X be integrable, then one may deduce directly from the differential equation (30) that a linear $\mathbb{E}[X \mid Y]$ implies a Gaussian X in this case too (see [12]). □

[3]It can also be shown that we necessarily have $\alpha_m < 0$ and m is odd, but these points are moot since we eventually have a contradiction.

Remark 1. The conclusion of Theorem 1 is derivable from the fact that $\mathbb{E}[X \mid Y = y] = O(y^2)$, shown in [15, Proposition 1.2] under the assumption that the input RV X has finite variance. Theorem 1 proves this conclusion under the more general setup when X is assumed to be only integrable.

III. Conditional Expectation Derivatives

We develop formulas for the higher-order derivatives of the conditional expectation, and establish upper bounds. The bounds in Theorem 2 on the norm of the derivatives of the conditional expectation will be crucial in Section IV for establishing a Bernstein approximation theorem that shows how well polynomials can approximate the conditional expectation in the mean-square sense.

Theorem 2. *Fix an integrable RV X and an independent $N \sim \mathcal{N}(0,1)$, and set $Y = X + N$. Let $r \ge 2$ be an integer, let C_r be as defined in (17), and denote $q_r := \lfloor(\sqrt{8r+9} - 3)/2\rfloor$ and $\gamma_r := (2rq_r)!^{1/(4q_r)}$. We have the bound*

$$\left\| \frac{d^{r-1}}{dy^{r-1}} \mathbb{E}[X \mid Y = y] \right\|_2 \le 2^r C_r \min\left(\gamma_r, \|X\|_{2rq_r}^r\right). \tag{39}$$

For $2 \le r \le 7$, we obtain the first few values of q_r as $1, 1, 1, 2, 2, 2$, and we have $q_r \sim \sqrt{2r}$ as $r \to \infty$ (see Remark 3 for a way to reduce q_r). To prove Theorem 2, we first express the derivatives of $y \mapsto \mathbb{E}[X \mid Y = y]$ as polynomials in the moments of the RV $X_y - \mathbb{E}[X_y]$, where X_y denotes the RV obtained from X by conditioning on $\{Y = y\}$.

Proposition 1. *Fix an integrable RV X and an independent $N \sim \mathcal{N}(0,1)$, and let $Y = X + N$. For each $(y,k) \in \mathbb{R} \times \mathbb{N}$, denote $f(y) := \mathbb{E}[X \mid Y = y]$ and*

$$g_k(y) := \mathbb{E}\left[(X - \mathbb{E}[X \mid Y])^k \mid Y = y\right]. \tag{40}$$

For $(\lambda_2, \cdots, \lambda_\ell) = \boldsymbol{\lambda} \in \mathbb{N}^*$, denote $\boldsymbol{g}^{\boldsymbol{\lambda}} := \prod_{i=2}^{\ell} g_i^{\lambda_i}$, with the understanding that $g_i^0 = 1$. Then, for every integer $r \ge 2$, we have that

$$f^{(r-1)} = \sum_{\boldsymbol{\lambda}\in\Pi_r} e_{\boldsymbol{\lambda}} \boldsymbol{g}^{\boldsymbol{\lambda}}, \tag{41}$$

where the integers $e_{\boldsymbol{\lambda}}$ are as defined in (15)-(16).

Proof Sketch. Differentiating the conditional expectation, we get $f' = g_2$. Further, differentiating g_k for $k \ge 1$ we obtain

$$g_k' = g_{k+1} - k g_2 g_{k-1}. \tag{42}$$

By differentiating both sides of $f' = g_2$ repeatedly and utilizing (42) after each differentiation step, we inductively obtain that $f^{(r-1)}$ admits a formula of the form

$$f^{(r-1)} = \sum_{\boldsymbol{\lambda}\in\Pi_r} h_{\boldsymbol{\lambda}} \boldsymbol{g}^{\boldsymbol{\lambda}} \tag{43}$$

for integers $h_{\boldsymbol{\lambda}}$. A closer look at how (43) is deduced from $f' = g_2$ and (42) reveals that the $h_{\boldsymbol{\lambda}}$ satisfy a recurrence relation of the form: for $\boldsymbol{\nu} \in \Pi_{r+1}$ there are integers $\{d_{\boldsymbol{\lambda},\boldsymbol{\nu}}\}_{\boldsymbol{\lambda}\in\Pi_r}$ such that

$$h_{\boldsymbol{\nu}} = \sum_{\boldsymbol{\lambda}\in\Pi_r} d_{\boldsymbol{\lambda},\boldsymbol{\nu}} h_{\boldsymbol{\lambda}}, \quad h_{(1)} = 1. \tag{44}$$

Further, the same recurrence in (44) generates the sequence $e_{\boldsymbol{\lambda}}$ defined in (16), thereby yielding $h_{\boldsymbol{\lambda}} = e_{\boldsymbol{\lambda}}$ for every $\boldsymbol{\lambda}$. □

Remark 2. The formula in Proposition 1 was derived in parallel to this work in [16]. Further, it is shown in [16] that formula (41) is the expansion of the r-th cumulant of X_y in terms of the moments of X_y via Bell polynomials.

Now we are ready to prove Theorem 2.

Proof of Theorem 2. We use the notation of Proposition 1. Fix $(\lambda_2, \cdots, \lambda_\ell) = \boldsymbol{\lambda} \in \Pi_r$. By the generalization of Hölder's inequality stating $\|\psi_1 \cdots \psi_k\|_1 \le \prod_{i=1}^k \|\psi_i\|_k$, we have that

$$\left\|\boldsymbol{g}^{\boldsymbol{\lambda}}(Y)\right\|_2^2 = \left\|\prod_{\lambda_i\neq 0} g_i^{2\lambda_i}(Y)\right\|_1 \le \prod_{\lambda_i\neq 0} \left\|g_i^{2\lambda_i}(Y)\right\|_s \tag{45}$$

where s is the number of nonzero entries in $\boldsymbol{\lambda}$. By Jensen's inequality for conditional expectation, for each i

$$\left\|g_i^{2\lambda_i}(Y)\right\|_s \le \|X - \mathbb{E}[X \mid Y]\|_{2i\lambda_i s}^{2i\lambda_i}. \tag{46}$$

Now, $r = \sum_{i=2}^{\ell} i\lambda_i \ge \sum_{i=2}^{s+1} i = \frac{(s+1)(s+2)}{2} - 1$, so we have that $s^2 + 3s - 2r \le 0$, i.e., $s \le q_r$. Further, $i\lambda_i \le r$ for each i. Hence, monotonicity of norms and inequalities (45) and (46) imply the uniform (in $\boldsymbol{\lambda}$) bound

$$\left\|\boldsymbol{g}^{\boldsymbol{\lambda}}(Y)\right\|_2 \le \|X - \mathbb{E}[X \mid Y]\|_{2rq_r}^{r}. \tag{47}$$

Observe that $\|X - \mathbb{E}[X \mid Y]\|_k \le 2\min\left((k!)^{1/(2k)}, \|X\|_k\right)$ (see [1]), so applying the triangle inequality in (41) we obtain

$$\left\|f^{(r-1)}(Y)\right\|_2 \le \sum_{\boldsymbol{\lambda}\in\Pi_r} c_{\boldsymbol{\lambda}} \left\|\boldsymbol{g}^{\boldsymbol{\lambda}}(Y)\right\|_2 \tag{48}$$

$$\le 2^r C_r \ \min\left(\gamma_r, \|X\|_{2rq_r}^{r}\right), \tag{49}$$

where $\gamma_r = (2rq_r)!^{1/(4q_r)}$, as desired. □

Remark 3. A closer analysis reveals that $i\lambda_i s$ in (46) cannot exceed $\beta_r := t_r^2(t_r + 1/2)$ for $t_r := (\sqrt{6r+7} - 1)/3$. For $r \to \infty$, we have $rq_r/\beta_r \sim 3^{3/2}/2 \approx 2.6$. The reduction when, e.g., $r = 7$, is from $rq_r = 14$ to $\beta_r = 10$.

IV. A Bernstein Approximation Theorem

We show that, if $X \sim p \in \mathscr{D}$ (see Definition 1), then the approximation error $\|E_n[X \mid Y] - \mathbb{E}[X \mid Y]\|_2$ decays faster than any polynomial in n.

Theorem 3. *Fix $p \in \mathscr{D}$, let $X \sim p$, suppose $N \sim \mathcal{N}(0,1)$ is independent of X, and set $Y = X + N$. There exists a sequence $\{D(p,k)\}_{k\in\mathbb{N}}$ of constants such that for all integers $n \ge \max(k-1, 1)$ we have*

$$\|E_n[X \mid Y] - \mathbb{E}[X \mid Y]\|_2 \le \frac{D(p,k)}{n^{k/2}}. \tag{50}$$

The proof relies on results on the Bernstein approximation problem in weighted L^p spaces. In particular, we consider the Freud case, where the weight is of the form e^{-Q} for Q of polynomial growth, e.g., a Gaussian weight.

Definition 2 (Freud Weights)**.** A function $W : \mathbb{R} \to (0,\infty)$ is called a *Freud Weight*, and we write $W \in \mathscr{F}$, if it is of the form $W = e^{-Q}$ for $Q : \mathbb{R} \to \mathbb{R}$ satisfying:

1) Q is even,
2) Q is differentiable, and $Q'(y) > 0$ for $y > 0$,
3) $y \mapsto yQ'(y)$ is strictly increasing over $(0,\infty)$,
4) $yQ'(y) \to 0$ as $y \to 0^+$, and
5) there exist $\lambda, a, b, c > 1$ such that for every $y > c$

$$a \le \frac{Q'(\lambda y)}{Q'(y)} \le b. \tag{51}$$

The convolution of a weight in $\mathscr{D}$ with the Gaussian weight $\varphi(x) := e^{-x^2/2}/\sqrt{2\pi}$ is a Freud weight. This can be shown by noting that with $p_Y = e^{-Q}$ we have $Q'(y) = \mathbb{E}[N \mid Y = y]$. We state this as a theorem here, and include the proof in [12].

Theorem 4. *If $p \in \mathscr{D}$ and $X \sim p$, then the probability density function of $X + N$, for $N \sim \mathcal{N}(0,1)$ independent of X, is a Freud weight.*

To be able to state the theorem we borrow from the Bernstein approximation literature, we need first to define the Mhaskar–Rakhmanov–Saff number.

Definition 3. If $Q : \mathbb{R} \to \mathbb{R}$ satisfies conditions (2)–(4) in Definition 2, and if $yQ'(y) \to \infty$ as $y \to \infty$, then the *n-th Mhaskar–Rakhmanov–Saff number* $a_n(Q)$ *of* Q is defined as the unique positive root a_n of the equation

$$n = \frac{2}{\pi}\int_0^1 \frac{a_n t Q'(a_n t)}{\sqrt{1-t^2}}\, dt. \tag{52}$$

Remark 4. The condition $yQ'(y) \to \infty$ as $y \to \infty$ in Definition 3 is satisfied if e^{-Q} is a Freud weight.

For example, the weight $W(y) = e^{-y^2}$, for which $Q(y) = y^2$, has $a_n(Q) = \sqrt{n}$ because $\int_0^1 t^2/\sqrt{1-t^2}\, dt = \frac{\pi}{4}$. If $X \sim p \in \mathscr{D}$, say $\operatorname{supp}(p) \subset [-M, M]$, and $p_Y = e^{-Q}$ (where $N \sim \mathcal{N}(0,1)$ is independent of X, and $Y = X + N$), then

$$a_n(Q) \le \left(2M + \sqrt{2}\right)\sqrt{n}. \tag{53}$$

This follows straightforwardly from $Q'(y) = \mathbb{E}[N \mid Y = y]$ (see [12] for the details).

We apply the following Bernstein approximation theorem [4, Corollary 3.6] to prove Theorem 3.

Theorem 5. *Fix $W \in \mathscr{F}$, and let u be an r-times continuously differentiable function such that $u^{(r)}$ is absolutely continuous. Let $a_n = a_n(Q)$ where $W = e^{-Q}$, and fix $1 \le s \le \infty$. Then, for some constant $D(W,r,s)$ and every $n \ge \max(r-1, 1)$*

$$\inf_{q\in\mathscr{P}_n} \|(q-u)W\|_{L^s(\mathbb{R})} \le D(W,r,s)\left(\frac{a_n}{n}\right)^r \|u^{(r)}W\|_{L^s(\mathbb{R})}. \tag{54}$$

Proof sketch of Theorem 3. The theorem follows by choosing $u(y) = \mathbb{E}[X \mid Y = y]$, $s = 2$, and $W = \sqrt{p_Y}$ in Theorem 5, and recalling (11), (53), and Theorems 2 and 4. □

REFERENCES

[1] D. Guo, Y. Wu, S. S. Shitz, and S. Verdu, "Estimation in gaussian noise: Properties of the minimum mean-square error," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2371–2385, 2011.

[2] A. Dytso, H. V. Poor, and S. S. Shitz, "A general derivative identity for the conditional mean estimator in gaussian noise and some applications," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 1183–1188.

[3] W. Alghamdi and F. P. Calmon, "Mutual information as a function of moments," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 3122–3126.

[4] D. Lubinsky, "A survey of weighted polynomial approximation with exponential weights," *Surveys in Approximation Theory*, vol. 3, pp. 1–105, 2007.

[5] Z. Ditzian and V. Totik, *Moduli of Smoothness*. Springer New York, 1987.

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

[7] Y. Wu and S. Verdu, "Functional properties of minimum mean-square error and mutual information," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 1289–1301, 2012.

[8] G. Han and J. Song, "Extensions of the I-MMSE relationship to gaussian channels with feedback and memory," *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5422–5445, 2016.

[9] G. Reeves, V. Mayya, and A. Volfovsky, "The geometry of community detection via the MMSE matrix," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 400–404.

[10] Y. Wu and S. Verdu, "MMSE dimension," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 4857–4879, 2011.

[11] S. Cha and T. Moon, "Neural adaptive image denoiser," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2981–2985.

[12] W. Alghamdi and F. P. Calmon, "Polynomial approximations of conditional expectations in scalar gaussian channels," https://arxiv.org/abs/2102.05970, 2021.

[13] OEIS Foundation Inc. (2021), The On-Line Encyclopedia of Integer Sequences, http://oeis.org/A032181.

[14] W. Bryc, *The Normal Distribution*. Springer New York, 1995.

[15] M. Fozunbal, "On regret of parametric mismatch in minimum mean square error estimation," in *2010 IEEE International Symposium on Information Theory*, 2010, pp. 1408–1412.

[16] A. Dytso, H. V. Poor, and S. Shamai (Shitz), "A general derivative identity for the conditional mean estimator in gaussian noise and some applications," https://arxiv.org/abs/2104.01883, 2021.