2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9518063

Neural Network-based Estimation of the MMSE

Mario Diaz*, Peter Kairouz†, Jiachun Liao‡, and Lalitha Sankar‡
*Universidad Nacional Autónoma de México, mario.diaz@sigma.iimas.unam.mx
†Google AI, kairouz@google.com
‡Arizona State University, {jliao15, lsankar}@asu.edu

***Abstract*—The minimum mean-square error (MMSE) achievable by optimal estimation of a random variable S given another random variable T is of much interest in a variety of statistical contexts. Motivated by a growing interest in auditing machine learning models for unintended information leakage, we propose a neural network-based estimator of this MMSE. We derive a lower bound for the MMSE based on the proposed estimator and the Barron constant associated with the conditional expectation of S given T. Since the latter is typically unknown in practice, we derive a general bound for the Barron constant that produces order optimal estimates for canonical distribution models.**

I. Introduction

The minimum mean-square error (MMSE) achievable by optimal estimation of a random variable given another one plays a pivotal role in classical statistics and information theory (see, e.g., [1]). More recently, the MMSE has been proposed as an average measure of information leakage in [2], [3], among others. Furthermore, it is closely related to probability of error, which has been used in the context of privacy as a proxy for information leakage in [4]–[6] and references therein.

The success of contemporary machine learning models comes in part from their massive complexity. However, this complexity comes at a cost in terms of interpretability and explainability, which is crucial for privacy and fairness considerations. As a result, there is a growing interest in developing tools and techniques to audit machine learning models for unwanted information leakage (e.g., [7], [8]). Indeed, there is a growing literature on data-driven methods for estimating different information leakage measures [9]–[11].

Consider two random variables S and T. In this work, we propose a neural network-based estimator for the MMSE in estimating S given T. In machine learning terms, this estimator is the minimum empirical loss attained by a two-layer neural network under the squared loss function. Using classical large deviations results and a theorem by Barron [12] regarding the universal approximation capabilities of two-layer neural networks, we derive a lower bound for the MMSE based on the proposed estimator and the Barron constant associated with the conditional expectation of S given T. Since the latter is typically unknown in practice, we derive a general bound for the Barron constant that produces order optimal estimates for canonical distribution models. In particular, it provides order optimal estimates in situations where additive Gaussian post-processing is used. By the relation between MMSE and probability of error, our main results also translate into provable data-driven lower bounds for probability of error.

This paper is organized as follows. In Section II we introduce our proposed estimator, discuss about the relevance of MMSE as a measure of information leakage, and review Barron's theorem. We derive a lower bound for the MMSE upon the proposed estimator and the Barron constant in Section III. In Section IV we derive a general bound for the Barron constant which yields order optimal estimates in some settings described in Section V. We provide some concluding remarks in Section VI.

II. Problem Setting and Preliminaries

A. Minimum Mean Square Error

Given real random variables S and T, the minimum mean square error (MMSE) in estimating S given T is defined as

$$\mathrm{mmse}(S|T) := \inf_{h \text{ meas.}} \mathbb{E}\left[(S - h(T))^2\right], \tag{1}$$

where the infimum is taken over all (Borel) measurable functions $h : \mathbb{R} \to \mathbb{R}$. Indeed, the infimum is attained by the conditional expectation of S given T, i.e.,

$$\mathrm{mmse}(S|T) = \mathbb{E}\left[(S - \eta(T))^2\right], \tag{2}$$

where $\eta(T) \overset{\text{a.s.}}{=} \mathbb{E}[S|T]$. Observe that if $S \overset{\text{a.s.}}{=} h_0(T)$ for some function $h_0 : \mathbb{R} \to \mathbb{R}$, then $\mathrm{mmse}(S|T) = 0$; also, if S and T are independent, then the MMSE is maximal and $\mathrm{mmse}(S|T) = \mathbb{E}\left[(S - \mathbb{E}[S])^2\right]$.

B. Neural Network-based MMSE Estimator

Let $\mathcal{H}_k$ be the hypothesis class associated with a two-layer neural network of size k with activation function hyperbolic tangent[1] (tanh). More specifically, $\mathcal{H}_k$ is the set of all functions $h : \mathbb{R} \to \mathbb{R}$ of the form

$$h(x) = c_0 + \sum_{j=1}^{k} c_j \tanh(a_j x + b_j), \tag{3}$$

where $a_j, b_j, c_j \in \mathbb{R}$. In this work, we propose the following neural network-based estimator for the MMSE in (1). Given a sample $\mathcal{S}_n = \{(S_1, T_1), \ldots, (S_n, T_n)\}$, we define

$$\mathrm{mmse}_{k,n}(S|T) := \inf_{h \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} (S_i - h(T_i))^2. \tag{4}$$

Observe that, optimization matters aside, $\mathrm{mmse}_{k,n}(S|T)$ can be computed using the sample $\mathcal{S}_n$ and a device capable of implementing a two-layer neural network of size k.

[1] This activation function is taken for concreteness; the results that follow hold for more general activation functions.

Our goal is to provide (probabilistic) bounds of the form

$$\mathrm{mmse}_{k,n}(S|T) - \mathrm{mmse}(S|T) \leq \epsilon_{k,n}, \tag{5}$$

where $\epsilon_{k,n}$ is a positive number depending on the neural network size k and the sample size n. It is important to remark that our motivation to study one-sided bounds as presented in (5) comes from model auditing in machine learning, where $\mathrm{mmse}_{k,n}(S|T) - \epsilon_{k,n}$ serves as a lower bound for $\mathrm{mmse}(S|T)$ – a measure of average information leakage from S to T.

C. MMSE as a Measure of Information Leakage

We now present two natural interpretations of the MMSE in terms of privacy in machine learning.

Let $\mathcal{H}$ be the hypothesis class of all measurable functions $h : \mathbb{R} \to \mathbb{R}$ and let $\ell_2 : \mathcal{H} \times \mathbb{R}^2 \to \mathbb{R}$ be the squared loss, i.e.,

$$\ell_2(h, (t, s)) = (s - h(t))^2. \tag{6}$$

Recall the definition of the MMSE given in (1). It is immediate to verify that

$$\mathrm{mmse}(S|T) = \inf_{h \in \mathcal{H}} L(h), \tag{7}$$

where $L(h) = \mathbb{E}\left[\ell_2(h, (T, S))\right]$ is the expected loss of h. Thus, the MMSE could be interpreted as the minimum expected loss over the (maximal) hypothesis class of all measurable functions. In the context of privacy, it corresponds to the expected loss of an adversary interested in inferring S from T that is capable of implementing any (measurable) function, i.e., the expected loss of the strongest adversary in terms of model capacity. Thus, the one-sided bound in (5) implies that $\mathrm{mmse}_{k,n}(S|T) - \epsilon_{k,n}$ is a lower bound for the expected loss of such adversary. This is particularly relevant in the context of model auditing, where an auditor could evaluate $\mathrm{mmse}_{k,n}(S|T) - \epsilon_{k,n}$ in a data driven manner and use it as a sanity check for potential unintended information leakage.

We also note that MMSE could serve as a lower bound for probability of error in the context of binary classification. Namely, if $S \in \{\pm 1\}$, then

$$\begin{aligned} P_{\text{error}}(S|T) &:= \inf_{h \text{ meas.}} \mathbb{E}\left[\mathbb{1}_{S \neq h(T)}\right] & (8)\\ &\geq \inf_{h \text{ meas.}} \mathbb{E}\left[\frac{1}{4}(S - h(T))^2\right] & (9)\\ &= \frac{1}{4}\mathrm{mmse}(S|T). & (10)\end{aligned}$$

Thus, the one-sided bound in (5) provides a (data-driven) lower bound for the probability of error of S given T.

D. Barron's Theorem

We end this section with a fundamental result by Barron [12] that establishes, in a quantitative manner, the universal approximation capabilities of two-layer neural networks.

For a distribution P and $\alpha \geq 1$, the (P, α)-norm of a measurable function $f : \mathbb{R} \to \mathbb{R}$ is given by

$$\|f\|_{P,\alpha} := \left(\int_{\mathbb{R}} |f(x)|^\alpha \mathrm{d}P(x)\right)^{1/\alpha}. \tag{11}$$

Also, recall that for a function $f : \mathbb{R} \to \mathbb{C}$, its Fourier transform, say $\hat{f} : \mathbb{R} \to \mathbb{C}$, is defined as

$$\hat{f}(\omega) := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} f(x) e^{-\mathrm{i}\omega x} \mathrm{d}x. \tag{12}$$

The following proposition, due to Barron [12], establishes in a quantitative manner the universal approximating capabilities of two-layer neural networks. For $K \subseteq \mathbb{R}$, the diameter of K is defined as

$$\mathrm{dia}(K) = \sup_{x,y \in K} |x - y|. \tag{13}$$

Proposition 1 (Proposition 1, [12])**.** *Let $k \geq 1$ and P a probability distribution supported over a compact set $K \subseteq \mathbb{R}$. If $h : \mathbb{R} \to \mathbb{R}$ is a smooth function*[2]*, then there exists $h_k \in \mathcal{H}_k$ such that*

$$\|h - h_k\|_{P,2} \leq \frac{\mathrm{dia}(K) C_h}{\sqrt{k}}, \tag{14}$$

where C_h is the so-called Barron constant of h defined as

$$C_h = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} |\omega||\hat{h}(\omega)| \mathrm{d}\omega. \tag{15}$$

Furthermore, the coefficients in (3) *may be restricted to satisfy $c_0 = h(0)$ and $\sum_{i=1}^k |c_i| \leq \mathrm{dia}(K) C_h$.*

III. 2-Layer Neural Network-based MMSE Estimation

For the remainder of this paper, we focus on a classification setting where S is a binary attribute with $S \in \{\pm 1\}$. The next theorem establishes, under mild conditions, a quantitative version of the desired bound (5).

Theorem 1. *Let $f_\pm : \mathbb{R} \to \mathbb{R}$ be the conditional density of T given $S = \pm 1$. Assume that (a) the support of f_- and f_+ are contained in a compact set $K \subset \mathbb{R}$ and (b) the function f_-/f_+ extends smoothly to an open set containing K. For all $k, n \in \mathbb{N}$, if $\delta > 0$, then, with probability at least $1 - \delta$,*

$$\mathrm{mmse}_{k,n}(S|T) - \mathrm{mmse}(S|T) \leq C_0 \sqrt{\frac{\log(1/\delta)}{2n}} + \frac{C_1}{k} + \frac{C_2}{\sqrt{k}}, \tag{16}$$

where C_0, C_1, C_2 are constants independent of k and n.

Moreover, if $\eta : \mathbb{R} \to \mathbb{R}$ is any smooth function such that $\eta(t) = \mathbb{E}[S|T = t]$ for all $t \in K$, then $C_0 \leq (2 + \mathrm{dia}(K) C_\eta)^2$, $C_1 \leq (\mathrm{dia}(K) C_\eta)^2$ and $C_2 \leq 4\mathrm{dia}(K) C_\eta$ where C_η is the Barron constant of η.

Proof. For ease of notation, let

$$\Delta_{k,n} := \mathrm{mmse}_{k,n}(S|T) - \mathrm{mmse}(S|T). \tag{17}$$

As established in (7), $\mathrm{mmse}(S|T)$ can be expressed in terms of the expected loss L associated with the squared loss ℓ_2. Indeed, it is the minimum of L over the set $\mathcal{H}$ of all measurable functions. Similarly, observe that $\mathrm{mmse}_{k,n}(S|T)$, as defined

[2] Barron's theorem holds in greater generality. However, for the purpose of this paper the present formulation suffices. We refer the reader to [12] for further details.

in (4), can be expressed in terms of the empirical loss $L_{\mathcal{S}_n}$ associated with ℓ_2. Indeed, we have that

$$\mathrm{mmse}_{k,n}(S|T) = \inf_{h \in \mathcal{H}_k} L_{\mathcal{S}_n}(h), \tag{18}$$

where $\mathcal{H}_k$ is the of all functions of the form (3) and

$$L_{\mathcal{S}_n}(h) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ell_2(h, (T_i, S_i)). \tag{19}$$

Therefore, we can write $\Delta_{k,n}$ as

$$\Delta_{k,n} = \inf_{h \in \mathcal{H}_k} L_{\mathcal{S}_n}(h) - \inf_{h \in \mathcal{H}} L(h). \tag{20}$$

Recall that the infimum defining the MMSE is attained by the conditional expectation. In particular, we have that

$$\inf_{h \in \mathcal{H}} L(h) = L(h^*), \tag{21}$$

where h^* is the conditional expectation of S given T, i.e., $h^*(T) \overset{\text{a.s}}{=} \mathbb{E}[S|T]$. Observe that, for every $t \in K$,

$$h^*(t) = \mathbb{E}(S \mid T = t) = \frac{\lambda f_+(t) - \overline{\lambda} f_-(t)}{\lambda f_+(t) + \overline{\lambda} f_-(t)}, \tag{22}$$

where $\lambda = \mathbb{P}(S = 1)$ and $\overline{\lambda} = \mathbb{P}(S = -1)$. By assumptions (a) and (b), there exists an integrable smooth function $\eta : \mathbb{R} \to \mathbb{R}$ that extends h^*, i.e, $\eta(t) = h^*(t)$ for all $t \in K$. Since $\eta = h^*$ over K, which contains the support of the distribution of T,

$$L(h^*) = \mathbb{E}[(S - h^*(T))^2] = \mathbb{E}[(S - \eta(T))^2] = L(\eta). \tag{23}$$

In particular, we have that

$$\Delta_{k,n} = \inf_{h \in \mathcal{H}_k} L_{\mathcal{S}_n}(h) - L(\eta). \tag{24}$$

Barron's theorem (Prop. 1) implies that there exists $\eta_k \in \mathcal{H}_k$ such that

$$\|\eta_k - \eta\|_{P_T,2} \le \frac{\mathrm{dia}(K) C_\eta}{\sqrt{k}}, \tag{25}$$

with P_T the distribution of T. Moreover, if we let

$$\eta_k(t) = c_0 + \sum_{i=1}^{k} c_i \sigma(a_i t + b_i), \tag{26}$$

the coefficients $c_0, \ldots, c_k$ can be taken such that $c_0 = \eta(0)$ and $\sum_i |c_i| \le \mathrm{dia}(K) C_\eta$. Observe that (22) implies that $|\eta(t)| \le 1$ for all $t \in K$. In particular, $\|\eta\|_{P_T,2} \le 1$ and, by our choice for the coefficients in (26),

$$\|\eta_k\|_\infty \coloneqq \sup_{t \in \mathbb{R}} |\eta_k(t)| \le 1 + \mathrm{dia}(K) C_\eta. \tag{27}$$

Continuing with (24), observe that

$$\Delta_{k,n} \le L_{\mathcal{S}_n}(\eta_k) - L(\eta_k) + L(\eta_k) - L(\eta). \tag{28}$$

Observe that $(S - \eta_k(T))^2 \le (1 + \|\eta_k\|_\infty)^2$. By the bound in (27), a routine application of Hoeffding's inequality implies that, with probability at least $1 - \delta$,

$$L_{\mathcal{S}_n}(\eta_k) - L(\eta_k) \le (2 + \mathrm{dia}(K) C_\eta)^2 \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{29}$$

Under the squared loss ℓ_2, it could be proved that for any two functions $h_1, h_2 : \mathbb{R} \to \mathbb{R}$,

$$|L(h_2) - L(h_1)| \le \|h_2 - h_1\|(2 + 2\|h_1| + \|h_2 - h_1\|), \tag{30}$$

where the norms are $(P_T, 2)$-norms. Therefore, by plugging (25) in (30), we obtain that

$$|L(\eta_k) - L(\eta)| \le \frac{\mathrm{dia}(K) C_\eta}{\sqrt{k}} \left(4 + \frac{\mathrm{dia}(K) C_\eta}{\sqrt{k}}\right). \tag{31}$$

By plugging (29) and (31) in (28), the theorem follows. □

Although the conditions of the previous theorem might seem restrictive, they could be easily guaranteed by adding a small-variance noise to T and then truncating the result. Indeed, if $\gamma > 0$ and $Z \sim \mathcal{N}(0, 1)$ is independent of (S, T), then (S, T') satisfy the hypotheses of the previous theorem with T' being the truncation of $T + \gamma Z$ to the interval $[-r, r]$ for any $r > 0$.

The assumptions of Theorem 1 guarantee the existence of a function η as in the second part of the statement. Nonetheless, to the best of the authors' knowledge, it is unknown how to find the function η that produces the smallest C_η and thus the sharpest bound (16).

While the right hand side of (16) decreases when k increases, note that $\mathrm{mmse}_{k,n}(S|T)$ also decreases when k increases. In fact, if $k \ge 2n$, then a two-layer neural network with k neurons can memorize the entire sample $\mathcal{S}_n$, leading to a trivial lower bound for $\mathrm{mmse}(S|T)$ as $\mathrm{mmse}_{k,n}(S|T) = 0$. Furthermore, the optimization minimization defining $\mathrm{mmse}_{k,n}(S|T)$ becomes harder as k increases. Overall, this reveals that finding the k that produces the best bound in (16) is a non-trivial task.

IV. A General Bound for the Barron Constant

Theorem 1 establishes a bound for the difference between $\mathrm{mmse}_{k,n}(S|T)$ and $\mathrm{mmse}(S|T)$ that depends on the neural network size k, the sample size n and the Barron constant C_η of (a smooth extension of) the conditional expectation of S given T. Providing estimates for the latter quantity might be a challenging task for two reasons: (i) the conditional expectation of S given T depends on the distribution of S and T, which is typically unavailable in practice; (ii) the Barron constant C_η is defined in terms of the Fourier transform of η, which makes its computation unfeasible in most cases. The next theorem alleviates the second issue by providing a general upper bound for C_η based on the L^1-norms of η', η'' and η'''.

Theorem 2. *Let $\eta : \mathbb{R} \to \mathbb{R}$ be a differentiable function. If C_η exists and $\eta^{(j)} \in L^1(\mathbb{R})$ for $j = 1, 2, 3$, then*

$$C_\eta \le \frac{2\sqrt{2}}{\sqrt{\pi}} \left(1 + \log\left(\frac{\sqrt{\|\eta'\|_1 \|\eta'''\|_1}}{\|\eta''\|_1}\right)\right) \|\eta''\|_1. \tag{32}$$

Proof. For ease of notation, let $\kappa \coloneqq \eta'$. As pointed out by Barron [12], if C_η exists then

$$C_\eta \coloneqq \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} |\hat{\kappa}(\omega)| \mathrm{d}\omega, \tag{33}$$

Indeed, up to technical details, (33) follows from the formula $\widehat{h'}(\omega) = \mathrm{i}\omega\hat{h}(\omega)$. For $0 < \lambda_1 < \lambda_2$, we split the integral in (33) into three parts:

$$\mathrm{I} := \int_{-\lambda_1}^{\lambda_1} |\hat{\kappa}(\omega)| \mathrm{d}\omega, \tag{34}$$

$$\mathrm{II} := \left(\int_{\lambda_1}^{\lambda_2} + \int_{-\lambda_2}^{-\lambda_1} \right) |\hat{\kappa}(\omega)| \mathrm{d}\omega, \tag{35}$$

$$\mathrm{III} := \left(\int_{\lambda_2}^{\infty} + \int_{-\infty}^{-\lambda_2} \right) |\hat{\kappa}(\omega)| \mathrm{d}\omega. \tag{36}$$

First, observe that

$$\mathrm{I} \leq 2\|\hat{\kappa}\|_\infty \lambda_1 \leq 2\|\kappa\|_1 \lambda_1, \tag{37}$$

where we applied the inequality $\|\hat{h}\|_\infty \leq \|h\|_1$. Also, since $\widehat{h'}(\omega) = \mathrm{i}\omega\hat{h}(\omega)$ whenever $h, h' \in L^1(\mathbb{R})$, we have that

$$\mathrm{II} \leq \left(\int_{\lambda_1}^{\lambda_2} + \int_{-\lambda_2}^{-\lambda_1} \right) \frac{1}{|\omega|} |\widehat{\kappa'}(\omega)| \mathrm{d}\omega \tag{38}$$

$$= 2\|\widehat{\kappa'}\|_\infty \log\left(\frac{\lambda_2}{\lambda_1}\right) \tag{39}$$

$$\leq 2\|\kappa'\|_1 \log\left(\frac{\lambda_2}{\lambda_1}\right). \tag{40}$$

Similarly, since $\widehat{h''}(\omega) = (\mathrm{i}\omega)^2\hat{h}(\omega)$ whenever $h, h'' \in L^1(\mathbb{R})$,

$$\mathrm{III} \leq \left(\int_{\lambda_2}^{\infty} + \int_{-\infty}^{-\lambda_2} \right) \frac{1}{\omega^2} \left|\widehat{\kappa''}(\omega)\right| \mathrm{d}\omega \leq 2\frac{\|\kappa''\|_1}{\lambda_2}. \tag{41}$$

By plugging (37), (40) and (41) in (33), we obtain that

$$C_\eta \leq \sqrt{\frac{2}{\pi}} \left(\|\kappa\|_1 \lambda_1 + \|\kappa'\|_1 \log\left(\frac{\lambda_2}{\lambda_1}\right) + \frac{\|\kappa''\|_1}{\lambda_2} \right). \tag{42}$$

By taking $\lambda_1 = \frac{\|\kappa'\|_1}{\|\kappa\|_1}$ and $\lambda_2 = \frac{\|\kappa''\|_1}{\|\kappa'\|_1}$, the result follows. □

In the following, for ease of notation, we let $\kappa = \eta'$. With this notation, the previous theorem shows that C_η can be controlled by the L^1-norms of κ, κ' and κ''. Specifically, we have that

$$C_\eta \leq \frac{2\sqrt{2}}{\sqrt{\pi}} \left(1 + \log\left(\frac{\sqrt{\|\kappa\|_1 \|\kappa''\|_1}}{\|\kappa'\|_1} \right) \right) \|\kappa'\|_1. \tag{43}$$

The following lemma provides some useful expressions for κ and its first two derivatives. Recall that if $f_\pm : \mathbb{R} \to \mathbb{R}$ is the conditional density of T given $S = \pm 1$ and $\lambda = \mathbb{P}(S = 1)$, then the conditional expectation of S given T is given by[3]

$$\eta(t) = \frac{\lambda f_+(t) - \bar{\lambda} f_-(t)}{\lambda f_+(t) + \bar{\lambda} f_-(t)}, \tag{44}$$

where $\overline{\lambda} := 1 - \lambda = \mathbb{P}(S = -1)$.

Lemma 1. *If η is defined as in (44) and $\kappa = \eta'$, then*

$$\kappa = 2\frac{g'_+ g_- - g_+ g'_-}{(g_+ + g_-)^2}, \tag{45}$$

$$\kappa' = 2\frac{g''_+ g_- - g_+ g''_-}{(g_+ + g_-)^2} - 2\kappa \frac{g'_+ + g'_-}{g_+ + g_-}, \tag{46}$$

$$\kappa'' = 2\frac{g'''_+ g_- + g''_+ g'_- - g'_+ g''_- - g_+ g'''_-}{(g_+ + g_-)^2} \tag{47}$$

$$- 2\kappa \frac{g''_+ + g''_-}{g_+ + g_-} - 3\kappa' \frac{g'_+ + g'_-}{g_+ + g_-}, \tag{48}$$

where $g_+ = \lambda f_+$ and $g_- = \bar{\lambda} f_-$.

Proof. Equation (45) follows easily from the equalities $\kappa = \eta'$ and $\eta = \frac{g_+ - g_-}{g_+ + g_-}$. By the quotient rule $\left(\frac{p}{q}\right)' = \frac{p'}{q} - \frac{p}{q}\frac{q'}{q}$, (46) follows easily from (45). Using similar arguments, (47) follows from (46). □

V. Barron Constant under Additive Gaussian Post-Processing

Despite the explicit upper bound for C_η given in Theorem 2 and the expressions for the derivatives of η in Lemma 1, it is still challenging to obtain numerical upper bounds in practice as in most applications the conditional density of T given S is unknown. Nonetheless, in this section we show that if the variable T is post-processed by adding Gaussian noise, then it is possible to give an upper bound for C_η using only the strength of the noise and mild information about the support of T given S. Furthermore, we show that such a bound is order optimal in some canonical distribution models.

Such a post-processing situation can arise when a data curator needs to release T in order to achieve some utility, but cannot do so due to privacy concerns. In this situation, a common practice is to produce a *sanitized* version T which is apt to be released. A common mechanism to attain this goal is the so-called additive Gaussian mechanism which produce a sanitized variable T^σ given by

$$T^\sigma = T + \sigma Z, \tag{49}$$

where $\sigma > 0$ and Z is a standard Gaussian r.v. independent of T. If $f_\pm$ denotes the conditional density of T given $S = \pm 1$, then the conditional density of T^σ given $S = \pm 1$ is given by

$$f^\sigma_\pm = f_\pm * K_\sigma, \tag{50}$$

where $*$ denotes the convolution operator and, for every $t \in \mathbb{R}$,

$$K_\sigma(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-t^2/2\sigma^2}. \tag{51}$$

We are interested in finding upper bounds for C_{η^σ} where η^σ is the conditional expectation of S given T^σ, i.e.,

$$\eta^\sigma := \frac{\lambda f^\sigma_+ - \bar{\lambda} f^\sigma_-}{\lambda f^\sigma_+ + \bar{\lambda} f^\sigma_-}. \tag{52}$$

To gain some intuition about the behavior of the Barron constant under the effect of additive Gaussian post-processing,

[3] More precisely, η is a smooth extension of the conditional expectation of S given T. In this sense, we abuse of the notation an use $f_\pm$ to denote a suitable extension of the conditional density of T given $S = \pm 1$.

in the next proposition we compute C_{η^σ} in the extreme case where $T = \mu S$ for some $\mu > 0$.

Proposition 2. *Let $\mu, \sigma > 0$. If $T = \mu S$, then $C_{\eta^\sigma} = \dfrac{\mu}{\sigma^2}$.*

Proof. A direct computation shows that

$$\eta(t) = \frac{e^{-(t-\mu)^2/2\sigma^2} - e^{-(t+\mu)^2/2\sigma^2}}{e^{-(t-\mu)^2/2\sigma^2} + e^{-(t+\mu)^2/2\sigma^2}} = \tanh\left(\frac{\mu t}{\sigma^2}\right). \quad (53)$$

Recall that $\widehat{h'}(\omega) = \mathrm{i}\omega\hat{h}(\omega)$ and thus

$$C_\eta = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} |\widehat{\eta'}(\omega)| \mathrm{d}\omega. \quad (54)$$

Observe that $\eta'(t) = \dfrac{\mu}{\sigma^2}\mathrm{sech}(\mu t)^2$. Using contour integration, it can be verified that

$$\widehat{\mathrm{sech}^2}(\omega) = \sqrt{\frac{\pi}{2}}\omega \,\mathrm{csch}\left(\frac{\pi}{2}\omega\right), \quad (55)$$

which is non-negative for all $\omega \in \mathbb{R}$. By applying the Fourier inversion formula to (54), we have that $C_\eta = \dfrac{\mu}{\sigma^2}\mathrm{sech}^2(0)$. □

The next theorem provides an upper bound for C_{η^σ} under the assumption that the support of T is bounded. Its proof relies on careful estimates of η^σ and its derivatives. We refer the interested reader to the full version of this paper [13].

Theorem 3. *Let $f_\pm$ be the conditional density of T given S. If $\mathrm{Supp}(f_\pm) \subset [-r, r]$ for some $r > 0$, then, for every $\sigma > 0$,*

$$C_{\eta^\sigma} \leq A + B\frac{r^2 M_0}{\sigma^4}\left[1 + \log\left(\frac{r^2 M_0(M_2 + rM_1 + r^2)}{\sigma^8}\right)\right], \quad (56)$$

where A, B are numeric constants and

$$M_p := \int_{\mathbb{R}} |t|^p \frac{\lambda f_+^\sigma(t)}{\lambda f_+^\sigma(t) + \bar{\lambda} f_-^\sigma(t)} \frac{\bar{\lambda} f_-^\sigma(t)}{\lambda f_+^\sigma(t) + \bar{\lambda} f_-^\sigma(t)} \mathrm{d}t. \quad (57)$$

Furthermore, if $\sigma \geq \sqrt[4]{10er^2M_0}$, then

$$C_{\eta^\sigma} \leq B\frac{r^2 M_0}{\sigma^4}\left[1 + \log\left(\frac{M_2}{r^2 M_0} + \frac{M_1}{rM_0} + 1\right)\right]. \quad (58)$$

Note that the bound in the previous theorem only require knowledge of the moment-like quantities M_p, as defined in (57). As we show below, in some situations $M_p = O(\sigma^{2(1+p)})$ as $\sigma \to \infty$. Therefore, in the large noise regime ($\sigma \gg 1$),

$$C_{\eta^\sigma} \leq O\left(\frac{\log(\sigma)}{\sigma^2}\right). \quad (59)$$

In view of Proposition 2, we conclude that the previous bound is order optimal (up to logarithmic factors).

Below we also show that in some situations $M_p = O(1)$ as $\sigma \to 0^+$. Therefore, in the small noise regime ($\sigma \ll 1$),

$$C_{\eta^\sigma} \leq O\left(\frac{\log(1/\sigma)}{\sigma^4}\right). \quad (60)$$

It is unclear at the moment if this bound is order optimal. Nonetheless, it is worth to remark that this bound is by no means trivial. Observe that, as $\sigma \to 0^+$, η^σ converges pointwise to η which in principle might have an unbounded Barron constant.

We end this section providing an upper bound for the moment-like quantities M_p in the case where the supports of $f_\pm$ are well-separated by some margin.

Proposition 3. *Let M_p be the quantities defined in (57). If there exist $r > 0$ and $\gamma \in (0, r)$ such that $\mathrm{Supp}(f_+) \subset [\gamma, r]$ and $\mathrm{Supp}(f_-) \subset [-r, -\gamma]$, then, for every $\sigma > 0$,*

$$M_0 \leq 2r + \frac{\sigma^2}{2\gamma}\frac{\lambda^2 + \bar{\lambda}^2}{\lambda\bar{\lambda}}e^{-2\gamma r/\sigma^2}, \quad (61)$$

$$M_1 \leq 2r^2 + \left[\frac{\sigma^4}{(2\gamma)^2} + \frac{r\sigma^2}{2\gamma}\right]\frac{\lambda^2 + \bar{\lambda}^2}{\lambda\bar{\lambda}}e^{-2\gamma r/\sigma^2}, \quad (62)$$

$$M_2 \leq 2r^3 + \left[\frac{2\sigma^6}{(2\gamma)^3} + \frac{2r\sigma^4}{(2\gamma)^2} + \frac{r^2\sigma^2}{2\gamma}\right]\frac{\lambda^2 + \bar{\lambda}^2}{\lambda\bar{\lambda}}e^{-2\gamma r/\sigma^2}. \quad (63)$$

In particular, $M_p = O(\sigma^{2(1+p)})$ as $\sigma \to \infty$ and $M_p = O(1)$ as $\sigma \to 0^+$.

The proof of the previous proposition relies on explicit estimates of $f_\pm^\sigma$. We refer the interested reader to the full version of this paper [13].

Observe that the leading term of M_p captures the parameters r, γ and λ in a natural way. More specifically, the leading term of M_p increases when the support r increases, the margin γ decreases and the data imbalance $|\lambda - 1/2|$ increases. As such, Proposition 3 captures natural intuitions in a quantitative manner.

VI. Conclusion

In this paper, we proposed a neural network-based estimator for the MMSE in estimating a random variable S given another random variable T. Motivated by model auditing in machine learning, we derived a lower bound for the sought MMSE based on the proposed estimator and the Barron constant of the conditional expectation of S given T. Finding meaningful estimates for the Barron constant might be a challenging task since: (i) the conditional expectation of S given T is rarely available in practice, and (ii) it is defined in terms of the Fourier transform of this conditional expectation. To alleviate the second issue, we provided a general upper bound for the Barron constant based on the L^1-norm of the conditional expectation and its derivatives. Furthermore, we showed that one can circumvent the first issue in applications where additive Gaussian post-processing is used. In such applications, our bounds for the Barron constant are indeed order optimal.

Acknowledgements

The work of L. Sankar and J. Liao is supported in part by NSF grants CIF-1901243, CIF-1815361, and CIF-2007688. The work of M. Diaz was supported in part by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) under grant IA101021.

References

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

[2] S. Asoodeh, F. Alajaji, and T. Linder, "Privacy-aware MMSE estimation," in *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 1989–1993.

[3] F. du Pin Calmon, A. Makhdoumi, M. Médard, M. Varia, M. Christiansen, and K. R. Duffy, "Principal inertia components and applications," *IEEE Transactions on Information Theory*, vol. 63, no. 8, pp. 5011–5038, 2017.

[4] I. Issa and A. B. Wagner, "Operational definitions for some common information leakage metrics," in *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 769–773.

[5] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Estimation efficiency under privacy constraints," *IEEE Transactions on Information Theory*, vol. 65, no. 3, pp. 1512–1534, 2018.

[6] A. Nageswaran and P. Narayan, "Data privacy for a ρ-recoverable function," *IEEE Transactions on Information Theory*, vol. 65, no. 6, pp. 3470–3488, 2019.

[7] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18.

[8] C. Song and V. Shmatikov, "Auditing data provenance in text-generation models," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 196–206.

[9] Z. Goldfeld, K. Greenewald, and Y. Polyanskiy, "Estimating differential entropy under Gaussian convolutions," *arXiv preprint arXiv:1810.11589*, 2018.

[10] M. Diaz, H. Wang, F. P. Calmon, and L. Sankar, "On the robustness of information-theoretic privacy measures and mechanisms," *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 1949–1978, 2019.

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

[12] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 930–945, 1993.

[13] M. Diaz, P. Kairouz, and L. Sankar, "Lower bounds for the minimum mean-square error via neural network-based estimation," in preparation.