HOW MANY BITS FROM HOW MANY SENSORS? A TRADE-OFF IN DISTRIBUTED NEAREST-NEIGHBOR LEARNING

Stefano Marano[(1)] *Vincenzo Matta*[(1)] *Peter Willett*[(2)] *

[(1)] Department of Information & Electrical Engineering and Applied Mathematics, University of Salerno, Fisciano (SA), Italy
[(2)] Electrical & Computer Engineering Department, University of Connecticut, Storrs, CT, USA.

ABSTRACT

In one of his landmark papers, Cover established the fundamental scaling laws of learning with nearest-neighbor rules [1]. With the recent advances on *distributed* nearest-neighbor learning in sensor networks novel trade-offs arise, involving the faithfulness of message representation (quantization bits) and the number of delivered messages (transmitting sensors). This is the main theme of this paper.

1. DISTRIBUTED LEARNING MODEL

One of the earliest and most famous results of Nearest-Neighbor (NN) learning is Cover's *half-information* result, well summarized by the statement: "*...it can be said that at least half the information in the infinite training set is contained in the nearest neighbor*" [1]. With reference to estimation problems with a Mean-Square-Error (MSE) criterion, this can be compactly expressed by the following ratio:

$$\frac{\mathsf{mmse}}{\mathsf{mse}_1} = \frac{1}{2}, \tag{1}$$

where mmse denotes the minimum MSE that can be achieved with perfect knowledge of the underlying data distributions, and mse_1 is the limiting (as the training-set size goes to infinity) MSE achieved using the nearest-neighbor rule.

In the same paper [1], the result is generalized to the case of k-NN rules, and the pertinent MSE-ratio becomes:

$$\frac{\mathsf{mmse}}{\mathsf{mse}_k} = \frac{1}{1+\frac{1}{k}}, \tag{2}$$

showing how the limiting learning accuracy scales with the number of neighbors k.

The above results refer to a centralized setup, where the training set needed to build the NN regression function is entirely available to the inference performer, which has to estimate a response variable Y_0 based upon a sensed observation X_0. In many networked systems, however, there is an increasing demand for *distributed* inference schemes, such that it makes sense to ask whether and how these fundamental laws apply in decentralized contexts.

This kind of questions gave rise to the emerging paradigm of *distributed learning* [2–7]: here the training set is disseminated through a network of agents communicating quantized versions of their labels to a Fusion Center (FC), which then produces the final estimate. Within the distributed learning framework, a decentralized version of the classical nearest-neighbor regression rule has been proposed [5], where only the k sensors owning the k training data closest to the observation X_0 (the k nearest neighbors)[1] communicate their information to the fusion center, exploiting an *ad-hoc* access policy based upon an ordered transmission protocol [8]. In this paper, we refer to this distributed NN strategy and, following [5], we start by showing that the MSE for a quantized k-NN rule obeys, in the limit of large training set sizes:

$$\frac{\mathsf{mmse}}{\mathsf{mse}_k^{(q)}} = \frac{1}{1+\frac{1}{k}\left(1+\frac{D_b}{\mathsf{mmse}}\right)}, \tag{3}$$

where $\mathsf{mse}_k^{(q)}$ is the limiting (large training set) MSE achieved with the k-NN *quantized* strategy, and D_b is the squared distortion of the NN labels due to a b-bits quantization. The above formula highlights that a trade-off exists between the number of neighbors k and the number of quantization bits b. This aspect is clearly seen when one must put a constraint in terms of storage/communication budget. The kind of constraint we consider here is the overall quantizers' rate R of the k transmitting sensors, that is, of the k nearest neighbors employed for the final estimation, yielding:

$$R = k \cdot b. \tag{4}$$

The main focus of this paper is in maximizing the MSE-ratio in (3), subject to a constraint on the allowed expense R in (4), which amounts to answering the following question: *Is it better to finely quantize few neighbors, or to roughly represent many of them?* (Or is the answer more often a compromise?)

2. RELATED WORK AND MAIN CONTRIBUTIONS

Distribution-free (universal, nonparametric) statistical learning [9, 10] has a long history in many branches of engineering and applied science. This notwithstanding, the problem of *decentralized* statistical learning is definitely a less mature research field. The problem has been systematically addressed in [2–4], where the authors show for the first time that universally consistent regression is possible, even in a distributed

*Peter Willett was supported by ONR under contract N000014-13-1-0231.

[1] We wish to avoid confusion here. Hereafter, the terminology "nearest-neighbors" does not refer to any geographic/topologic attribute. It refers to closeness of the training data to the current measurement X_0.

sensor network, with reference to a decentralized implementation of the *naive kernel* estimator. Then, the result has been extended to the case of decentralized NN rules [5–7], using an access policy that relies upon the idea of ordered transmissions, originally proposed in [8]. Both the decentralized naive kernel and the decentralized NN make strong use of the randomized quantizers proposed in [11, 12], which are suited to universal and decentralized inference problems.

The present work stems from some computer-based experimental evidences found in [5], where it is observed that, in comparing the 1-NN strategy with infinite precision (number of bits $b \gg 1$), to the k-NN rule with one-bit quantized data, an interesting trade-off arises, in terms of quantizers' precision and active neighbors. Here we propose to exploit this trade-off, with reference to: i) a distributed k-NN rule; ii) randomized quantizers with resolution b; iii) a constraint on the overall quantization rate $R = k \cdot b$; and iv) an increasingly large training-set size.

3. DISTRIBUTED LEARNING MODEL

The statistical learning problem we are faced with can be formalized as follows. We want to estimate a response variable $Y_0 \in \mathbb{R}$, based on the sensed data $X_0 \in \mathbb{R}^d$, when the joint statistical distribution of the pair (X_0, Y_0) is *unknown* [9]. With reference to a *supervised* learning model, we assume the availability of a training set $T_n = \{(X_i, Y_i)\}_{i=1}^n$, that is a collection of independent, identically distributed (i.i.d.) realizations of (X_0, Y_0). An estimator of Y_0 is then represented by: $r_n : \mathbb{R}^d \longrightarrow \mathbb{R}$, where the *regression function* $r_n(x_0) = r_n(x_0, T_n)$ depends on x_0 and on the training set T_n. It is standard to omit the explicit dependence of r_n upon the training set for notational simplicity. As a performance proxy, we adopt the MSE, namely, the quantity $\mathbb{E}\{[r_n(X_0) - Y_0]^2\}$. By application of the orthogonality principle, one has:

$$\begin{aligned} &\mathbb{E}\{[r_n(X_0) - Y_0]^2\} \\ &\quad = \quad \mathsf{mmse} + \mathbb{E}\{[r_n(X_0) - r^*(X_0)]^2\}, \end{aligned} \tag{5}$$

where $r^*(x_0) = \mathbb{E}\{Y_0 | X_0 = x_0\}$ is the optimal estimator, also referred to as the (optimal) regression function.

A popular choice for the regression function is that based on the k-NN rule, and we here focus on that. Let $\big\{\big(X_{(i,n)}(x_0), Y_{(i,n)}(x_0)\big)\big\}_{i=1}^n$ be the sequence of pairs ordered according to

$$\|X_{(1,n)}(x_0) - x_0\| \leq \cdots \leq \|X_{(n,n)}(x_0) - x_0\|,$$

where $\|\cdot\|$ denotes the standard Euclidean norm in $\mathbb{R}^d$. We rule out ties by assuming continuous random variables[2]. The k-NN regression function we are interested in is accordingly

$$\frac{1}{k} \sum_{i=1}^{k} Y_{(i,n)}(x_0) = \sum_{i=1}^{n} W_{ni}(x_0)\, Y_i, \tag{6}$$

where the latter representation is convenient for later use, and is written in terms of the weights:

$$W_{ni}(x_0) = \begin{cases} 1/k, & \text{if } X_i \text{ is one of the } k\text{-NN of } x_0, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

[2]Even when this is not the case, the observation space can be artificially enlarged by including a random continuous component, as detailed in [9].

Note that the weights are functions of x_0 and of the observation variables in the training set $\{X_i\}_{i=1}^n$.

As anticipated, in this paper we deal with a *distributed* version of the above problem. Adhering to the standard model proposed in [2–4], we consider a training set disseminated through a network of n sensors that is deployed for estimation purposes: without loss of generality, assume that each sensor reads a single example (X_i, Y_i) from T_n. At a certain time, the observation variable X_0 is made available to the FC, and is broadcast to all nodes. By exploiting the locally available examples, sensors deliver messages to the FC, which produces the final estimate.

In order to implement the NN rule in a decentralized way, sensors transmit their labels with transmitting delays chosen as a function of the observed distances $\|X_i - X_0\|$, thus enabling the FC to read the labels already ordered according to the desired NN criterion. This access policy is based on the well-assessed strategy of *ordered* transmissions, proposed in [8], and then applied to different inference problems, see, e.g., [13, 14]. Recently, it has been tailored to decentralized NN rules in [5–7]. The specific details on the implementations of this access protocol can be found in the aforementioned works, and are here omitted for space limitations. For the subsequent discussion, it suffices to assume that the FC is able to receive the labels $Y_{(i,n)}(x_0)$, namely, those corresponding to the k nearest neighbors to x_0.

3.1. Randomized quantizers for universal estimation

The last sentence of the above section is deliberately wrong. Indeed, we would like to stress that, in our decentralized problem, one must be faced with some data quantization. And indeed, what we can say is that the FC recovers some *quantized* version of the labels, say $\mathcal{Q}_b(Y_{(i,n)}(x_0))$, where $\mathcal{Q}_b(y)$ is a b-bits quantizer of the input y.

The main difficulty here is that the distribution of Y is unknown. To overcome this issue, we resort to the *universal* randomized quantization rule proposed in [11, 12]. In order to quantize a value y to b bits, we proceed as follows. Assume from now on that Y is a bounded random variable, with $|Y| \leq V$. First, divide the range $[-V, V]$ into intervals of length $\Delta = 2V/(2^b - 1)$, yielding the thresholds $-V + i\Delta$, with $i = 0, 1, \ldots, 2^b - 1$. Then, find the interval where y lies, and denote by $\tau(y)$ the corresponding lower threshold. Finally, round y to one of the endpoints by a biased coin flip. Thus, *for a given* y, the quantizer output is a binary random variable taking values $\tau(y)$ and $\tau(y) + \Delta$. As shown in [11,12], setting $p_b(y) = \mathbb{P}\{\mathcal{Q}_b(y) = \tau(y) + \Delta\} = \frac{y - \tau(y)}{\Delta}$, makes the quantizers unbiased, i.e., $\mathbb{E}\{\mathcal{Q}_b(y)\} = y$, where the expectation operator acts w.r.t. to the quantizers' randomness only, since y is here fixed. Similarly:

$$\mathrm{VAR}\{\mathcal{Q}_b(y)\} = \frac{(2V)^2}{(2^b - 1)^2}\, p_b(y)\,[1 - p_b(y)]. \tag{8}$$

3.2. MSE computation with quantized labels

The unquantized k-NN regression function in (6) can be modified in order to take into account the quantization of the labels Y_i, yielding the following regression function that can be

computed at the FC:

$$r_n(x_0) = \sum_{i=1}^{n} W_{ni}(x_0)\mathcal{Q}_b(Y_i). \tag{9}$$

It is useful to remark that $i)$ in addition to the training set, $r_n(x_0)$ contains a further source of randomness, due to the adoption of *probabilistic* quantizers; and that $ii)$ this fact does not impair the orthogonality principle (5), such that it suffices to focus on the regression error $\mathbb{E}\{[r_n(X_0) - r^*(X_0)]^2\}$.

By defining $Q_i = \mathcal{Q}_b(Y_i)$, it is expedient to reinterpret the last term in (5) as the k-NN regression function for estimating a fictitious parameter $Q_0 = \mathcal{Q}_b(Y_0)$, based upon the observation X_0 and the modified training set $\{X_i, Q_i\}_{i=1}^n$. The optimal regression function for this problem is

$$\mathbb{E}\{Q_0|X_0 = x_0\} = \mathbb{E}\{Y_0|X_0 = x_0\} = r^*(x_0), \tag{10}$$

which is nothing but the optimal regression function corresponding to the original (X, Y) learning problem. The above follows from the fact that, given Y_0 and X_0, the residual randomness is in the quantizers' output, which are unbiased in this conditional space. From Problem 6.4 in [9] it then follows that:

$$\mathbb{E}\{[r_n(X_0) - r^*(X_0)]^2\} \overset{n\to\infty}{\longrightarrow} \frac{\mathbb{E}\{[r^*(X_0) - Q_0]^2\}}{k}. \tag{11}$$

Invoking again the conditional unbiasedness of the quantizers, the last quantity can be written as:

$$\frac{1}{k}\left(\underbrace{\mathbb{E}\{[r^*(X_0) - Y_0]^2\}}_{\mathsf{mmse}} + \underbrace{\mathbb{E}\{[Q_0 - Y_0]^2\}}_{D_b}\right). \tag{12}$$

Using now (11) and (12) into (5), the limiting MSE with quantized data is:

$$\mathsf{mse}_k^{(q)} = \mathsf{mmse} + \frac{1}{k}\left(\mathsf{mmse} + D_b\right),$$

which corresponds to the MSE-ratio in (3). Furthermore, by application of (8), the quantization distortion D_b can be written as:

$$D_b = \frac{(2V)^2}{(2^b-1)^2}\,\mathbb{E}\{p_b(Y)[1 - p_b(Y)]\}. \tag{13}$$

The adopted universal setup prevents us from a precise evaluation of D_b, since this quantity depends[3] on the unknown distribution of Y via the term $\mathbb{E}\{p_b(Y)[1 - p_b(Y)]\}$. For this reason we next focus on the worst-case bound $D_b \leq \frac{V^2}{(2^b-1)^2}$, holding because $p(1-p) \leq 1/4$ for $p \in [0, 1]$. This yields the lower bound for the MSE-ratio:

$$\frac{\mathsf{mmse}}{\mathsf{mse}_k^{(q)}} \geq \frac{1}{1 + \frac{1}{k} + \frac{1}{k}\frac{(2^b-1)^{-2}}{\overline{\mathsf{mmse}}}} = \mathcal{J}(\overline{\mathsf{mmse}}, b, k), \tag{14}$$

where $\overline{\mathsf{mmse}} = \mathsf{mmse}/V^2$ is the scaled mmse. It is worth noting that the lower bound $\mathcal{J}$ depends on the unknown (X, Y)-distribution only through the scaled mmse, which will be a key property to be exploited in the next section.

[3]Note that the distortion depends only on the marginal (prior) distribution of the parameter Y, not on the statistical relationship between X and Y.

4. OPTIMIZATION PROBLEM

Let us introduce the set $\mathcal{D}(R) = \{b, k \in \mathbb{N} :\ b \cdot k \leq R\}$. We are interested in maximizing the MSE-ratio (lower bound), for a given available budget R. We start by formulating the problem when the mmse is known. The optimized ratio is formally given by:

$$\mathcal{J}^*(\overline{\mathsf{mmse}}, R) = \max_{(b,k)\in\mathcal{D}(R)} \mathcal{J}(\overline{\mathsf{mmse}}, b, k). \tag{15}$$

Since, however, the *actual* $\overline{\mathsf{mmse}}$ depends upon the underlying distribution and is therefore generally unknown, we resort to the following max-min approach. Observe first that $\overline{\mathsf{mmse}} \leq 1$, since $\mathsf{mmse} \leq \mathbb{E}\{Y^2\}$ because 0 is a legitimate estimator, and $Y^2 \leq V^2$ by our boundedness assumption. Then, in a sense, the ratio $\overline{\mathsf{mmse}}$ quantifies a *relative* accuracy with respect to the range of the estimated variable. In practical problems, it is reasonable to restrict the analysis to a class where this scaled error cannot be lower than a given value, call it ϵ. Accordingly, we formulate the max-min problem:

$$\max_{(b,k)\in\mathcal{D}(R)} \min_{\overline{\mathsf{mmse}}\geq\epsilon} \mathcal{J}(\overline{\mathsf{mmse}}, b, k) = \mathcal{J}^*(\epsilon, R), \tag{16}$$

where the last equality easily follows from the monotonicity property of $\mathcal{J}(\overline{\mathsf{mmse}}, b, k)$ with respect to $\overline{\mathsf{mmse}}$.

4.1. Approximate closed-form solution

Equation (16) reveals that it is useful to understand the main properties of the optimized lower bound $\mathcal{J}^*(\epsilon, R)$, as a function of ϵ. Accordingly, we now study in more detail the optimization problem in (15). This is amenable to a direct numerical solution, and we shall pursue this approach in the next section. Before doing that, we would like to provide some approximations that turn out to be useful from a theoretical standpoint. A closer look to (14) reveals that: $i)$ for small values of $\overline{\mathsf{mmse}}$, the quantizer resolution is expected to be increased, and $ii)$ due to the exponential dependence of the quantization error upon b, for values of R that are not too small, the optimal unconstrained value of b is seldom expected to exceed the available R. As a result, we propose to replace b with a continuous counterpart β, and to solve the unconstrained, continuous optimization problem

$$\beta^*(\overline{\mathsf{mmse}}) = \frac{1}{R}\,\arg\min_{\beta\in\mathbb{R}} \underbrace{\beta\left(1 + \frac{(2^\beta - 1)^{-2}}{\overline{\mathsf{mmse}}}\right)}_{f(\beta)}, \tag{17}$$

having used $k \approx R/\beta$, and considering only the non-constant terms at the denominator of $\mathcal{J}$ in (14). The optimal $\beta^*(\overline{\mathsf{mmse}})$ solves $f'(\beta) = 0$, where:

$$f'(\beta) = 1 + \frac{1}{\overline{\mathsf{mmse}}}\left[(2^\beta - 1)^{-2} - (2\beta\ln 2)\,2^\beta(2^\beta - 1)^{-3}\right].$$

This task has to be performed numerically, as we show in the next section. In addition, we would like to give a closed-form approximation, holding in the high resolution regime, where we can write:

$$f'(\beta) \approx 1 + \frac{1}{\overline{\mathsf{mmse}}}2^{-2\beta}\left[1 - 2\beta\ln 2\right].$$

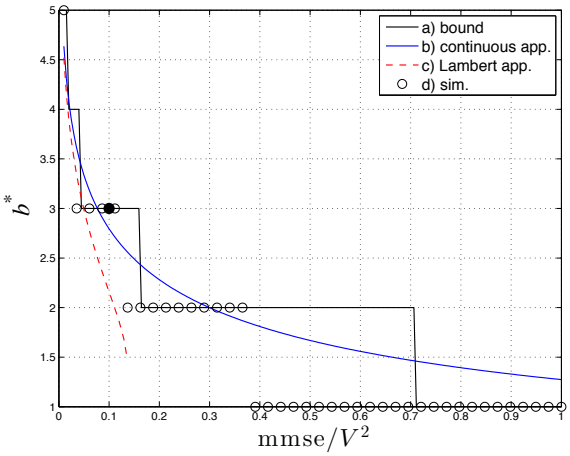

Fig. 1. Optimal number of bits, as a function of $\overline{\mathsf{mmse}} = \mathsf{mmse}/V^2$, for cases $a) - d)$ as described in the main text. The black-filled point represents the optimal number of bits solving (16), with $\epsilon = 0.1$.

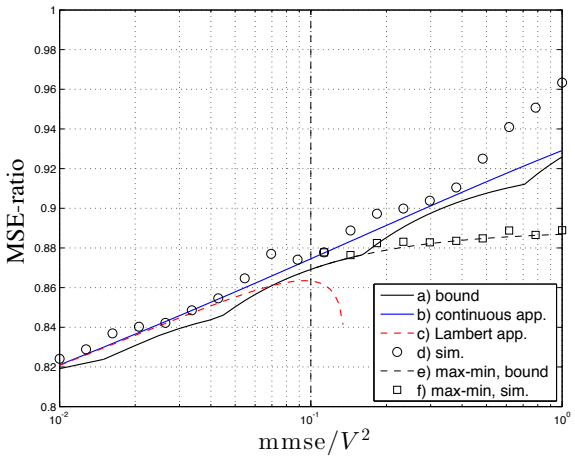

Fig. 2. MSE-ratios, as a function of the scaled minimum error $\overline{\mathsf{mmse}} = \mathsf{mmse}/V^2$, for cases $a) - f)$ as described in the main text. The dashed vertical line corresponds to the worst-case error $\epsilon = 0.1$ used in solving (16).

Setting $w = 1 - 2\beta \ln 2$, and equating to zero, we have

$$w\, e^w = -e \cdot \overline{\mathsf{mmse}} \Rightarrow w = \mathcal{W}_{-1}(-e \cdot \overline{\mathsf{mmse}}),$$

yielding:

$$\beta^*(\overline{\mathsf{mmse}}) \approx \frac{1 - \mathcal{W}_{-1}(-e \cdot \overline{\mathsf{mmse}})}{2 \ln 2}, \qquad (18)$$

where $\mathcal{W}_{-1}(x)$, with $x \in (-1/e, 0)$, denotes the lower[4] branch of the Lambert function. Note that the condition $x \in (-1/e, 0)$ implies that the above approximation can be used provided that, at least, $\overline{\mathsf{mmse}} < 1/e^2$.

5. RESULTS AND SUMMARY

Let us now construct some illustrative examples aimed at testing the presented results. Without loss of generality, we work with $V = 1$. Let $U(a_1, a_2)$ be a random variable uniform in the interval $[a_1, a_2]$, and let[5]

$$X = U(-1/2, 1/2), \qquad Y = r^*(X) + \mathcal{E}, \qquad (19)$$

with optimal regression function $r^*(x) = a/(1+a)\sin(2\pi x)$. For the range $\overline{\mathsf{mmse}} < 1/3$, we use the error model

$$\mathcal{E} = \frac{U(-1,1)}{1+a} \Leftrightarrow \overline{\mathsf{mmse}} = \mathbb{E}\{\mathcal{E}^2\} = \frac{1}{3}\frac{1}{(1+a)^2}, \qquad (20)$$

such that, when a ranges from ∞ to 0, $\overline{\mathsf{mmse}}$ ranges from 0 to $1/3$. To explore the complementary interval $(1/3, 1)$, we consider the alternative model, for $a \in (0, 1)$:

$$\mathcal{E} = \mathcal{B}\frac{U(-1, -1+a)}{1+a} + (1-\mathcal{B})\frac{U(1-a, 1)}{1+a}, \qquad (21)$$

where $\mathcal{B}$ is an equiprobable binary random variable. It can be shown that $\overline{\mathsf{mmse}} = \frac{1-a+a^2/3}{(1+a)^2}$, which allows spanning the range $(\frac{1}{12}, 1)$. We set the constraint to $R = 25$. In Fig. 1 we show the optimized number of bits computed with different methodologies, namely: $a)$ maximization of the lower bound in (15); $b)$ continuous unconstrained optimization (17); $c)$ the limiting approximation based upon the Lambert function (18); and $d)$ maximization of the actual MSE-ratio for the introduced model, evaluated by 10^5 Monte Carlo runs. With reference to $a)$, we see that the general trend is as follows: for small values of $\overline{\mathsf{mmse}}$, the optimal recipe goes in the direction of using finely quantized observations, for instance, with $\overline{\mathsf{mmse}} = 10^{-2}$ in the figure we get $b^* = k^* = 5$. As $\overline{\mathsf{mmse}}$ increases, we go in the opposite direction, reaching, for $\overline{\mathsf{mmse}} = 1$, the situation $b^* = 1$, $k^* = R = 25$, yielding $\mathcal{J}^* = (1 + 2/R)^{-1}$. The approximations of this solution, see $b)$ and $c)$, seem to be satisfying, in that $b)$ is a good smoothed version of the actual, staircase function corresponding to $a)$, and $c)$ tends to $b)$ as $\overline{\mathsf{mmse}}$ decreases. Finally, the curve in $d)$, pertaining to the specific models described by (19), (20) and (21), exhibits the same kind of behavior, while, of course, the actual values are not equal (and they must not be), since we have optimized a bound. The corresponding situation is terms of MSE-ratio is displayed in Fig. 2, where it is seen that the actual performance is effectively lower bounded by the theoretical $\mathcal{J}^*$, and that the overall behavior is even more faithful, due to the fact that the discretization issues are expected to be less pronounced when embedded in the overall MSE performance.

The above considerations refer to the behavior of the various optimized MSE-ratios as function of the actual mmse, which is a preliminary step to address the worst-case analysis in (16), which is addressed by curves $e)$ and $f)$, with reference to $\epsilon = 0.1$. Specifically, $e)$ shows the bound $\mathcal{J}(\overline{\mathsf{mmse}}, b, k)$, when b and k are chosen so as to optimize the worst-case as in (16); and $f)$ displays the actual MSE-ratio for the considered data model, for the same values of b and k. As it must be, the considered design is conservative, in the sense that the MSE-ratios, in the region $\overline{\mathsf{mmse}} > \epsilon$, are lower than the ones optimized with knowledge of the true $\overline{\mathsf{mmse}}$, while, in the point $\overline{\mathsf{mmse}} = \epsilon$ (i.e., the worst case in the considered region), the MSE-ratio is in fact maximized.

[4]The lower branch corresponds to $w \leq -1$, which is of interest here because we are working in the high resolution regime, and $w = 1 - 2\beta \ln 2$.

[5]Note that a regression problem can be always written as $Y = r^*(X) + \mathcal{E}$, where $\mathcal{E} = Y - r^*(X)$, and $\mathbb{E}\{\mathcal{E}^2\} = \mathsf{mmse}$.

6. REFERENCES

[1] T. M. Cover, "Estimation by the nearest neighbor rule," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 1, pp. 50–55, Jan. 1968.

[2] J. B. Predd, S. R. Kulkarni, and H. V. Poor, "Distributed learning in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 56–69, Jul. 2006.

[3] ——, "A collaborative training algorithm for distributed learning," *IEEE Trans. Inf. Theory*, vol. 55, no. 4, pp. 1856–1871, Apr. 2009.

[4] ——, "Consistency in models for distributed learning under communication constraints," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 52–63, Jan. 2006.

[5] S. Marano, V. Matta, and P. Willett, "Nearest-neighbor distributed learning by ordered transmissions," *IEEE Trans. Signal Process.*, vol. 61, no. 21, pp. 5217–5230, Nov. 2013.

[6] S. Marano, V. Matta, and P. Willett, "Nearest-neighbor distributed learning under communication constraints," in *Proc. IEEE ICASSP*, Vancouver, Canada, May 2013, pp. 3278–3282.

[7] S. Marano, V. Matta, and P. Willett, "Decentralized nearest-neighbor learning over noisy channels: the uncoded way," in *Proc. 16th International Conference on Information Fusion, FUSION 2013*, Istanbul, Turkey, July 2013, pp. 426–431.

[8] R. S. Blum and B. M. Sadler, "Energy efficient signal detection in sensor networks using ordered transmissions," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3229–3235, Jul. 2008.

[9] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer, 2002.

[10] C. J. Stone, "Consistent nonparametric regression," *The Annals of Statistics*, vol. 5, no. 4, pp. 595–620, Jul. 1977.

[11] Z.-Q. Luo, "Universal decentralized estimation in a bandwidth constrained sensor networks," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 2210–2219, Jun. 2005.

[12] J.-J. Xiao, S. Cui, Z.-Q. Luo, and A. J. Goldsmith, "Power scheduling of universal decentralized estimation in sensor networks," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 413–422, Feb. 2006.

[13] R. S. Blum, "Ordering for estimation and optimization in energy efficient sensor networks," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2847–2856, Jun. 2011.

[14] P. Braca, S. Marano, and V. Matta, "Single-transmission distributed detection via order statistics," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 2042–2048, Apr. 2012.