NEAREST-NEIGHBOR DISTRIBUTED LEARNING UNDER COMMUNICATION CONSTRAINTS

Stefano Marano, Vincenzo Matta

DIEII, University of Salerno,
via Ponte don Melillo I-84084,
Fisciano (SA), Italy.
E-mails: {marano, vmatta}@unisa.it.

*Peter Willett**

ECE Department,
U-2157, University of Connecticut,
Storrs, CT 06269 USA.
E-mail: willett@engr.uconn.edu.

ABSTRACT

A wireless sensor network is engaged in a statistical learning task, to be accomplished in a decentralized fashion. The focus here is in *distributed* Nearest-Neighbor (NN) regression, in the presence of communication constraints. We first introduce a general channel access policy which allows the fusion center to recover training-set labels ordered according to the NN criterion, in the absence of any data exchange among sensors. Then, two different paradigms are considered, where the communication cost is measured as: i) the channel accesses; ii) the quantization bits. In the former scenario, we propose a distributed NN strategy reaching an asymptotic performance of twice the minimum achievable mean-square error, with *only one sensor* transmitting information. In the latter case, we achieve *universally consistent* distributed NN regression even with one-bit quantized labels.

Index Terms— Statistical learning, Nonparametric inference, Universal estimation, Distributed learning, Wireless sensor networks.

1. DISTRIBUTED LEARNING MODEL

Distribution-free or nonparametric inference [1] is a well-established discipline, nowadays ubiquitous in a number of real-world applications. It basically deals with estimating a response variable $Y_0 \in \mathbb{R}$, based on a measured observation variable $X_0 \in \mathbb{R}^d$, when the joint statistical distribution of (X_0, Y_0) is *completely unknown* [1]. In the specific context of *supervised* learning, meaningful estimation is made possible by the availability of a training set $T_n = \{(X_i, Y_i)\}_{i=1}^n$, namely, a collection of independent, identically distributed (i.i.d.) copies of (X_0, Y_0).

An estimator of Y_0 can be formally represented as:

$$r_n : \mathbb{R}^d \longrightarrow \mathbb{R},$$

where $r_n(x_0) = r_n(x_0, T_n)$ is a function of x_0 and of the training set T_n, the explicit dependence upon this latter being usually omitted for notational simplicity. One of the classical performance measure is the Mean Square Error (MSE), which, by the orthogonality principle, can be written as:

$$\mathbb{E}\{[r_n(X_0) - Y_0]^2\} = \mathbb{E}\{[r_n(X_0) - r^*(X_0)]^2\} + \text{MMSE}, \quad (1)$$

where $r^*(x_0) = \mathbb{E}[Y_0|X_0 = x_0]$ is the Minimum Mean Square Error (MMSE) estimator, also called the (optimal) regression function. A common practice is to look for *universally consistent* estimators, i.e., the ones reaching the optimal regression function as n gets large (consistency), irrespective of the underlying distribution (universality). Since the second term at the RHS in eq. (1) is independent from the particular estimation strategy, the goodness of any estimator $r_n(x_0)$ can be simply measured by the L_2 error with respect to the optimal regression function, leading to the following definition of consistency:

DEFINITION 1 [1, def. 1.3] *A sequence of regression function estimates $\{r_n\}$ is called weakly universally consistent if*

$$\lim_{n\to\infty} \mathbb{E}\left\{[r_n(X_0) - r^*(X_0)]^2\right\} = 0$$

for all distributions of (X, Y) with $\mathbb{E}[Y^2] < \infty$.

This paper deals with a *decentralized* version of the above problem, following the emerging paradigm of distributed learning proposed in [2–4]. A network of n sensors is deployed for estimation purposes, and the training set is disseminated through it: to make things simple, and without loss of generality, each sensor reads a single example (X_i, Y_i) from T_n. At a certain epoch, the observation variable X_0 is made available to the Fusion Center (FC), which broadcasts it to all nodes. By exploiting the locally available examples, sensors deliver messages to the FC, which produces the final estimate.

The network topology is parallel, and remote units have limited complexity and available energy, implying severe communication constraints on the channels from sensors to the FC. Conversely, the reverse link is essentially unconstrained, such that X_0 is perfectly recovered by the nodes.

*P. Willett was supported by the U.S. Office of Naval Research under Grant N00014-10-10412.

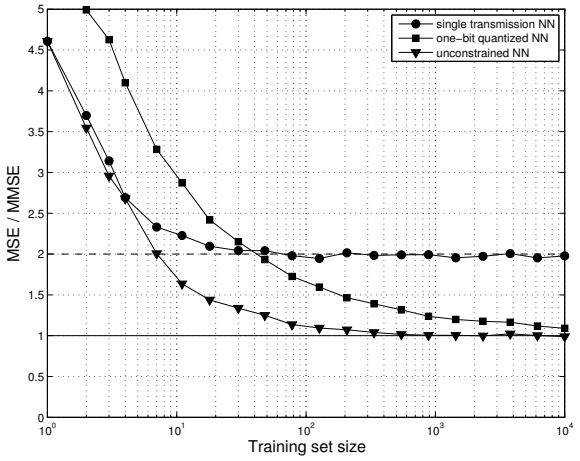

Fig. 1. MSE versus training set size for the example in eq. (2), with MMSE=1/4. Note that MSE is normalized to optimum (MMSE).

2. RELATED WORK AND MAIN CONTRIBUTIONS

Despite the long history of nonparametric regression, the problem of distributed learning under communication constraints has been systematically addressed only recently [2–4]. In [4] the authors show for the first time that *distributed* universally consistent regression is possible, by using the properties of the probabilistic universal quantizers proposed in [5, 6]. More specifically, they propose a decentralized version of the classical *naive kernel* estimator where the sensors have three possibility: to send 0, 1 or to abstain. Moreover, in [4] it is noted that "*...nearest neighbor rules do not apply as a given agent's decision rule would then need to depend on the data observed by the other agents; such interagent communication is not allowed in the current model*".

In this manuscript we try to overcome this difficulty by exploiting a peculiar access policy based on ordered transmissions, which allows us implementing Nearest Neighbor (NN) rules in a fully decentralized fashion. The idea of ordered transmissions (in the context of *parametric* detection) has been originally proposed in [7], and further developed in [8, 9]. Capitalizing on this access policy, we propose two distributed NN schemes, suited to different communication constraints:

i) The communication cost is measured in terms of channel accesses. A scheme with *only one* sensor transmitting to the FC is proposed, exhibitng an asymptotic performance of twice the MMSE.

ii) The communication cost is measured in terms of quantization bits. A strategy with local quantizers is designed, and proved to be universally weakly consistent, regardless of the quantizers' resolution.

As a preview of the results obtained in this paper, we consider, without any pretence of generality, the following simple example[1]:

$$X \sim \mathcal{U}(-1,1), \quad Y = \Lambda(X) + \sqrt{\text{MMSE}}\, W, \tag{2}$$

where $\mathcal{U}(-1,1)$ is an uniform random variable with support $[-1,1]$; $\Lambda(x)$ is a triangular wave of period 2, with unitary peak amplitude and $\Lambda(0) = 0$; W is a standard Gaussian, independent of X. The optimal regression function is accordingly $r^*(x_0) = \Lambda(x_0)$.

The relative merits between the two proposed schemes are shown in Fig. 1. Notably, for scheme i) the asymptotic limit of $2\times$ MMSE is practically met with just a dozen of elements in the training set, and with only one sensor effectively sending information. On the other hand, scheme ii) with one-bit quantizers needs some more examples to get going, but it reaches the unbeatable MMSE. For comparison purposes, the classical NN algorithm corresponding to the bandwidth-unconstrained system is also reported.

3. NEAREST-NEIGHBOR ACCESS

A classical choice for building estimators relies upon the so-called *local averaging* regression functions [1]:

$$\sum_{i=1}^{n} W_{ni}(x_0)\, Y_i, \tag{3}$$

where the weights are functions of x_0 and of the observation variables in the training set, namely, $X_1, X_2, \ldots, X_n$. For the NN regression, the weights are written as:

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

In order to conceive a *decentralized* implementation of a local averaging regression function, two properties are key. First, in eq. (3) the observation variables X_i act only on the weights $W_{ni}(x_0)$, so that they are basically decoupled from the response variables Y_i. In addition, the weights of the NN rule exhibit an on-off structure. This suggests as a fundamental design guideline the following decoupled approach: an *access policy* aimed at reproducing the weights at the FC, where only the sensors with nonzero weights transmit their labels; a *coding* strategy acting only on these relevant labels Y_i to be transmitted.

In the mentioned case of the naive kernel, this idea translates into the scheme of distributed regression with abstention proposed in [4]. A further simplification there is that the weight corresponding to the i-th sensor depends only on the i-th observation (and on X_0), such that abstention can be completely determined by a *locally available* knowledge. In our case of NN regression this fails to be true. With the general "decoupling" philosophy in mind, we now introduce a different access policy specifically tailored to the NN-rules.

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

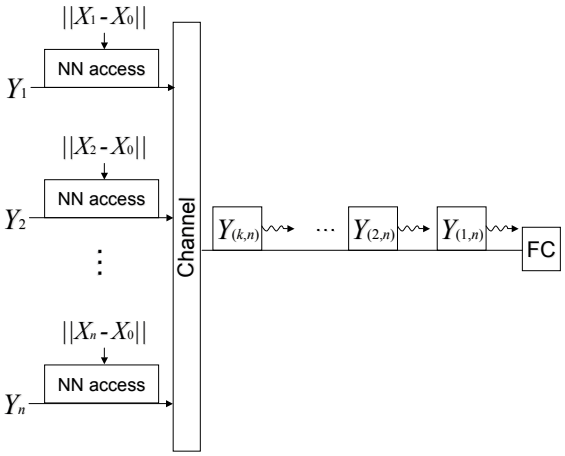

Fig. 2. NN access. The i-th sensor will transmit its label Y_i at a time instant based upon the distance $||X_i - X_0||$. This allows the FC to recover the labels ordered according to the NN criterion. Here $Y_{(i,n)}$ is a shortcut for $Y_{(i,n)}(X_0)$.

Let:

$$\left(X_{(1,n)}(x_0), Y_{(1,n)}(x_0)\right), \ldots, \left(X_{(n,n)}(x_0), Y_{(n,n)}(x_0)\right),$$

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 NN-regression function we are interested is accordingly $(1/k)\sum_{i=1}^{k} Y_{(i,n)}(x_0)$.

The proposed NN access policy is as follows, see Fig. 2. The i-th sensor evaluates the *local* distance $\|X_i - X_0\|$ between its training observation X_i, and the current X_0. Then, it sends the label Y_i over the channel, at a transmission instant directly related (say, proportional) to $\|X_i - X_0\|$. The origin of the transmission time axis can be safely set when the sensors receive a certain broadcast message from the FC, for instance, when the common observation X_0 is sent. It is worth noting that the distances $\|X_{(i,n)} - X_0\|$ vanish as n grows, so that in practice the proportionality constant must properly scale with n to avoid collapse. Assuming that each sensor is informed about the value of n, the scaling law for the proportionality constant can be chosen by resorting to the powerful theory of spacings [10]. Details on this are here omitted for space reasons, and will be reported elsewhere [11].

With this protocol, the nodes with smaller values $\|X_i - X_0\|$ transmit first: the FC gets the labels ordered just in terms of the desired NN criterion. Thus, the FC is able to compute the weights relevant for the evaluation of (3), or, equivalently, the ordered labels $Y_{(i,n)}(x_0)$. Note that this is obtained without transmitting the value of the observation variables X_i.

[2]Even when this is not the case, the observation space can be artificially enlarged by including a random continuous component, as detailed in [1].

4. SINGLE-TRANSMISSION LEARNING

Under this paradigm we impose the communication cost in terms of channel accesses. This philosophy has a well-established tradition in the literature (see, *e.g.*, [12]), and is suited to those applications where the burden associated to higher-layer protocols is significant, such that, when communication takes place, it convenient to send high-resolution data, here taken as continuous random variables.

We are specifically interested in the most economic operational point of a single transmission. The proposed strategy can be schematically summarized:
1) each sensor prepares the *unquantized* label Y_i, to be sent according to the NN access policy;
2) after receiving the first (quickest) delivery, the FC inhibits any further transmission by a broadcast stop message.
Therefore, the FC collects only the nearest-neighbor label $Y_{(1,n)}(X_0)$. Using the nearest neighbor for estimation purposes has a long history, since the pioneering work by Cover [13]. Using the known results in [13], we can state (without proof) the following claim.

PROPOSITION *(Single-transmission Nearest-Neighbor). Let $Y_{1,n}(X_0)$ be the nearest-neighbor label received by the fusion center with the single-transmission protocol. Let $r^*(x)$ be the optimal regression function, and define the conditional variance $\sigma^2(x) = \mathbb{E}[Y^2|X = x] - r^*(x)^2$. Under the assumption of finite second moment for X, if there exist positive constants A and B, such that, for any $x_1, x_2 \in \mathbb{R}^d$:*

$$\begin{aligned} [r^*(x_1) - r^*(x_2)]^2 &\leq A\|x_1 - x_2\|^2 \\ |\sigma^2(x_1) - \sigma^2(x_2)| &\leq B\|x_1 - x_2\|^2, \end{aligned}$$

then, for $n \to \infty$,

$$\mathbb{E}\left\{[Y_{1,n}(X_0) - Y_0]^2\right\} \longrightarrow 2\,\text{MMSE},$$

that is, asymptotically, a loss of a factor 2 is suffered, with respect to the optimal MMSE *estimator.* □

Remarkably, the above proposition shows that we can build an universal, distributed NN estimator by using *only one transmitting sensor*, with the performance of twice the MMSE.

5. UNIVERSALLY CONSISTENT NN RULES WITH UNIVERSAL QUANTIZATION

This can be perhaps considered as a more conventional communication paradigm. Here:
1) each sensor prepares a *quantized* version of the label Y_i, again to be sent according to our NN transmission policy;
2) the FC inhibits any further transmission by a broadcast stop message, after having received the first k_n deliveries.

The lack of knowledge about the underlying distribution prevent us from classical quantizer design. We instead need some *universal* quantization rule. We accordingly resort to

the *probabilistic universal* quantizers proposed in [5, 6]. Assume first that Y is a bounded random variable, with $|Y| \leq V$, and that we want to quantize it into b bits. We accordingly divide $[-V, V]$ into intervals of length $\Delta = (2V)/(2^b - 1)$, obtaining the thresholds $\tau_i = -V + i\Delta$, $i = 0, 1, \ldots, 2^b - 1$. Then, we find the interval where y lies, and round it to one of the endpoints by a biased coin flip. For instance, if $y \in [\tau_i, \tau_{i+1})$: $\mathbb{P}\{\mathcal{Q}(y; b) = \tau_{i+1}\} = p$. The output of the quantizer is thus a binary random variable taking values τ_{i+1} and τ_i, with probability p and $1 - p$, respectively. For any fixed y, the above quantizers can be made *unbiased*, a property that is key in building consistent estimators. As a matter of fact, the choice $p = (y - \tau_i)/\Delta$ yields $\mathbb{E}\{\mathcal{Q}(y; b)\} = y$ (expectation computed w.r.t. to the randomness of the quantizers only, y being deterministic). The quantizer variance is easily evaluated and upper bounded as $\mathbb{E}\{[\mathcal{Q}(y; b) - y]^2\} \leq \Delta^2/4$, having used the relation $p(1 - p) \leq 1/4$, for $p \in [0, 1]$.

The above arguments hold if Y is bounded. This limitation is easily removed as follows, see [4]. For an *unbounded* Y, arbitrarily choose a range $[-V, V]$, where the quantizer works as described earlier. When $|y| > V$ choose uniformly at random between $\pm V$. Thus, whenever $|y| > V$, $\mathcal{Q}(y; b)$ is a zero-mean binary random variable taking values $\pm V$, with variance V^2. Furthermore, we shall allow the quantizers' support to depend upon the number of sensors n, namely $V = V_n$. With an abuse of notation, we accordingly will use $\mathcal{Q}_n(y) = \mathcal{Q}(y; b, V_n)$, where the dependence on V_n is contained in the suffix n, and that on b is suppressed, because we shall work with an arbitrary, but fixed, number of bits. We finally have

$$\mathbb{E}\{\mathcal{Q}_n(y)\} = y\, I_{\mathcal{V}_n}(y), \qquad \text{VAR}\{\mathcal{Q}_n(y)\} \leq V_n^2, \tag{4}$$

where $\mathcal{V}_n = [-V_n, V_n]$, and $I_{\mathcal{V}_n}(y)$ is the indicator function of the set $\mathcal{V}_n$.

We are now ready to present the distributed NN algorithm with quantized labels. As said, a label Y_i must be quantized before being transmitted according to the NN access rule. Accordingly, the FC builds an estimated regression function close in shape to (3), with the original labels replaced with their quantized counterparts:

$$r_n(x_0) = \sum_{i=1}^{n} W_{ni}(x_0)\, \mathcal{Q}_n(Y_i). \tag{5}$$

We have the following

THEOREM *(Quantized k_n-NN). For any $k_n \to \infty$, with $k_n/n \to 0$, the estimated regression function (5) is weakly universally consistent for all distributions of (X, Y) with $\mathbb{E}\{Y^2\} < \infty$, provided that $V_n \to \infty$, and that $V_n^2/k_n \to 0$.* □

Sketch of proof. We denote by $\bar{r}_n(x_0)$ the expectation of $r_n(x_0)$ conditioned on T_n and X_0:

$$\bar{r}_n(x_0) = r_n^{NN}(x_0) - \sum_{i=1}^{n} W_{ni}(x_0) Y_i\, I_{\mathcal{V}_n^c}(Y_i), \tag{6}$$

where $\mathcal{V}_n^c$ is the complement of $\mathcal{V}_n$ with respect to $\mathbb{R}$, and $r_n^{NN}(x_0) = (1/k_n) \sum_{i=1}^{k_n} Y_{(i,n)}(x_0)$ is the desired NN regression function with unquantized data. The above follows from the fact that, conditioned on T_n and X_0, the only randomness is in the quantizers' output, whose average is computed by the first of eqs. (4). The undesired error term ascribed to the overload region of the quantizers can be controlled by letting the range V_n to diverge in a suitable way, as we shall promptly show. We can write

$$\begin{aligned} \mathbb{E}\left\{[r_n(X_0) - r^*(X_0)]^2\right\} &= \mathbb{E}\left\{[r_n(X_0) - \bar{r}_n(X_0)]^2\right\} \\ &+ \mathbb{E}\left\{[\bar{r}_n(X_0) - r^*(X_0)]^2\right\}, \end{aligned}$$

which is upper bounded by

$$\begin{aligned} &2\, \underbrace{\mathbb{E}\{[r_n^{NN}(X_0) - r^*(X_0)]^2\}}_{\text{classical } k_n\text{-NN } L_2 \text{ error}} + 2\, \underbrace{\mathbb{E}\{[\sum_{i=1}^{n} W_{ni}(X_0) Y_i\, I_{\mathcal{V}_n^c}(Y_i)]^2\}}_{\text{overload error}} \\ &+ \mathbb{E}\{\underbrace{\mathbb{E}\{[r_n(X_0) - \bar{r}_n(X_0)]^2 | T_n, X_0\}}_{\text{conditional variance of } r_n(X_0)}\}, \end{aligned} \tag{7}$$

having used eq. (6) and the well-known sum-of-squares inequality $(\sum_{j=1}^{k} a_j)^2 \leq k \sum_{j=1}^{k} a_j^2$. The first term is the difference between the optimal regression function $r^*(X_0)$ and the classical k_n-NN estimator, which is known to reach weak consistency [1], provided that $k_n \to \infty$ and $k_n/n \to 0$. As to the overload term, using the sum-of-squares inequality and assumption 1 of Stone's theorem [1, 14], it is upper bounded, for a given $c > 0$, by $c\, \mathbb{E}\{Y^2 I_{\mathcal{V}_n^c}(Y)\}$, which clearly vanishes as n goes to infinity. Finally, using the independence among the quantizer randomizations, the inner expectation in the last term of eq. (7) can be rewritten as

$$\sum_{i=1}^{n} W_{ni}^2(X_0) \text{VAR}\{\mathcal{Q}_n(Y) | T_n, X_0\} \leq \frac{V_n^2}{k_n}.$$

The inequality follows by using the second of eqs. (4). By assumption $V_n^2/k_n \to 0$, which ends the proof. •

6. CONCLUSIONS

Distributed nearest-neighbor regression under different communication constraints is addressed: when the constraint is on the channel accesses, a $2\times$ MMSE performance is reached with *only one sensor* transmitting; when the bit-rate matters, the optimal MMSE is achievable, even with *one-bit* quantizers. This summarizes the main technical findings of the paper.

Beyond the opening example discussed in Sect. 2, space reasons prevent us from providing computer experiments corroborating our convergence results and, more important, showing the convergence *rates* of the different schemes. A more in-depth investigation of these issues will be reported elsewere. We stress, finally, that the estimation schemes presented in this work emerge as the basic building blocks for distributed learning over the wireless channel, where the effect of noisy and fading links must be incorporated [11].

7. REFERENCES

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

[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] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] R. Pyke, "Spacings," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 27, no. 3, pp. 395–449, 1965.

[11] S. Marano, V. Matta, and P. Willett, "Nearest-Neighbor Distributed Learning over Communication Channels," *IEEE Trans. Signal Process.*, *submitted*, 2013.

[12] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: a low-communication-rate scheme for distributed detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, no. 2, pp. 554–568, Apr. 1996.

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

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