Data-aided Sensing where Communication and Sensing Meet: An Introduction

Jinho Choi
School of Information Technology
Deakin University, Australia
(e-mail: jinho.choi@deakin.edu.au).

Abstract—Since there are a number of Internet-of-Things (IoT) applications that need to collect data sets from a large number of sensors or devices in real-time, sensing and communication need to be integrated for efficient uploading from devices. In this paper, we introduce the notion of data-aided sensing (DAS) where a base station (BS) utilizes a subset of data that is already uploaded and available to select the next device for efficient data collection or sensing. Thus, using DAS, certain tasks in IoT applications, including federated learning, can be completed by uploading from a small number of selected devices. Two different types of DAS are considered: one is centralized DAS and the other is distributed DAS. In centralized DAS, the BS decides the uploading order, while each device can decide when to upload its own local data set among multiple uploading rounds in distributed DAS. In distributed DAS, random access is employed where the access probability of each device is decided according to its local measurement for efficient uploading.

Index Terms—Data-aided sensing; Internet-of-Things (IoT); Cross-Layer; Federated Learning

I. INTRODUCTION

It is expected that the Internet of Things (IoT) will play an important role in a number of applications in Industry 4.0, including smart cities and factories, in the future [1] [2]. There could be a number of IoT systems and it is desirable to build them as layered systems with interoperability, where the bottom layer is responsible for collecting and processing information or data from devices or sensors [3].

For IoT connectivity, a number of solutions are studied including WiFi, cellular IoT, and so on. Compared to others, cellular IoT can support IoT applications over a large area. For example, as discussed in [4], narrowband IoT (NB-IoT) [5] can be used to support IoT applications over a large area through cellular systems. In cellular IoT (for the bottom layer in IoT systems), each base station (BS) can be used as a data collector from devices or sensors deployed over a cell. Since long-term evolution (LTE) BSs are well deployed, cellular IoT might play a crucial role as IoT infrastructure in collecting a large amount of data from devices including mobile phones and sensors.

In a number of IoT applications, collecting data sets from devices deployed in an area requires devices’ sensing to acquire local measurements or data and uploading to a server. In cellular IoT, devices are to send their local measurements to a BS. While sensing and uploading can be considered separately, they can also be integrated, which leads to data-aided sensing (DAS) [6]. In general, DAS can be seen as iterative data collection scheme where a BS is to collect data sets from devices or nodes1 through multiple rounds. In DAS, the BS chooses a set of nodes in each round based on the data sets that are already available at the BS from previous rounds for efficient data collection. As a result, the BS (actually a server that is connected to the BS) is able to efficiently provide an answer to a given query with a small number of measurements compared to random polling.

In this paper, we introduce the key idea and approaches of DAS [6] [7] [8]. In particular, two different DAS schemes, namely centralized DAS and distributed DAS, are discussed. Depending on applications, in DAS, different objective functions can be used in selecting the devices in each round. We consider an entropy-based objective function to illustrate the idea of DAS and also discuss another objective function when measurements have a sparse representation [6], where the notion of compressive sensing (CS) [9] [10] is exploited.

While the BS decides uploading order in centralized DAS, each user can decide whether or not to upload its local data in each round in distributed DAS. Due to uncoordinated transmissions in distributed DAS, random access is employed with parallel multiple channels, where the access (or uploading) probability of each device can be decided by its measurement. This approach can be applied to federated learning [11] [12] [13], as studied in [8].

It is noteworthy that this paper is to introduce DAS that has been recently studied in [6] [7] [8]. As a result, this paper can be seen as a review paper, while it is new to classify DAS into centralized and distributed ones and introduce them.

Notation: Matrices and vectors are denoted by upper- and lower-case boldface letters, respectively. The superscript $T$ denotes the transpose. For a set, $\mathcal{X}$, its complement set is denoted by $\mathcal{X}^c$. $\mathbb{E}[]$ and $\text{Var}(\cdot)$ denote the statistical expectation and variance, respectively. In addition, $\text{Cov}(\cdot)$ represents the covariance matrix of random vector $\mathbf{x}$. $\mathcal{CN}(\mathbf{a}, \mathbf{R})$ and $\mathcal{CN}(\mathbf{a})$ represent the distributions of real-valued Gaussian and circularly symmetric complex Gaussian (CSCG) random vectors with mean vector $\mathbf{a}$ and covariance matrix $\mathbf{R}$, respectively.

II. SYSTEM MODEL

Suppose that there are a number of devices (say $K$ devices) within a cell and each device has a local measurement or data

1 Throughout the paper, it is assumed that nodes, devices, and sensors are interchangeable.
to upload to a BS that is to collect a data set of measurements. For example, each device can be seen as a sensor node that is deployed over a certain area to collect local environmental data, or a mobile device that is associated with a certain application that needs to learn from a set of data from a number of mobile devices (e.g., federated learning [11] [12] [13]). When the bandwidth is limited with a large \( K \), it is difficult to simultaneously upload all the measurements from devices. Thus, there should be multiple transmissions or uploading rounds to be carried out as illustrated in Fig. 1.

The objective function in (1) varies depending on the application. As an example, let us consider an entropy-based DAS where the BS is to choose the uploading order based on the entropy of measurements. Since \( \mathcal{X}(t) \) is available at the BS after iteration \( t \), the entropy or information of the remained measurements becomes

\[
H(\mathcal{X}^c(t) | \mathcal{X}(t)),
\]

where \( H(X | Y) \) represents the conditional entropy of \( X \) for given \( Y \). Thus, the objective function can be given as

\[
\Theta(x_k, \mathcal{X}(t)) = H(\mathcal{X}^c(t) | \mathcal{X}(t)) - H(x_k | \mathcal{X}(t)),
\]

where \( H(x_k | \mathcal{X}(t)) \) is the amount of information by uploading the measurement from device \( k \). Clearly, for fast data collection (or data collection with a small number of devices), we want to choose device \( k \) to minimize the objective function.

Note that \( H(\mathcal{X}^c(t) | \mathcal{X}(t)) \) is independent of \( k \). Thus, the next device is to be chosen according to the maximization of conditional entropy is given by

\[
k(t + 1) = \arg\max_{k \in \mathcal{K}(t)} H(x_k | \mathcal{X}(t)).
\]

That is, the next device should have the maximum amount of information (in terms of the entropy) given that \( \mathcal{X}(t) \) is already available at the BS.

For an illustration, suppose that there are \( K = 20 \) devices or nodes, which are deployed in a unit square, with their local measurements. Assume that the correlation of measurements is proportional to the distance between devices, i.e., the correlation coefficient is \( \rho_{i,k} = e^{-||u_i - u_k||} \), where \( u_k \) denotes the coordination of device \( k \). In Fig. 2, the locations of 20 devices are represented by square markers, while the first device, \( k(1) \), and the last device, \( k(K) \), are represented by star and cross markers, respectively. In addition, each edge represents the link between nodes \( k(t+1) \) and \( k(t) \). In general, there is no edge between two adjacent devices due to a high correlation, and the next device tends to be far away from the current device to minimize the entropy gap (as a far device likely has a highly uncorrelated measurement with the current device’s one).

In [7], the mean squared error (MSE) between the total measurement, \( x = [x_1 \ldots x_K]^T \), and its estimate \( \hat{x}_k(x_k, \mathcal{X}(t-1)) \), which is an estimate of \( x \) from \( x_k \) and \( \mathcal{X}(t-1) \), is to be minimized in choosing the device in round \( t \), i.e.,

\[
\Theta(x_k, \mathcal{X}(t)) = \mathbb{E} \left[ ||x - \hat{x}_k(x_k, \mathcal{X}(t-1))||^2 | \mathcal{X}(t-1) \right],
\]

where the expectation is carried out over \( \mathcal{X}^c(t-1) \), which is the subset of measurements that are not available before round \( t \). Since the measurements are assumed to be Gaussian in [7], second order statistics can be used for DAS, which is called Gaussian DAS.
shows reconstruction errors at the BS with DAS and RRS is known at the BS. To be precise, let $x$ be the signal vector obtained from all the nodes’ measurements, i.e., $x$, has a sparse representation [14] [15] [16]. Let $[x]_k = x_k \in \mathbb{R}, k \in K = \{1, \ldots, K\}$. Furthermore, $x$ can be represented by a sparse vector $s \in \mathbb{R}^{M \times 1}$. To this end, it is assumed that

$$x_k = b_k^T s, \quad k \in K,$$

where $b_k \in \mathbb{R}^{M \times 1}$ is the measurement vector at node $k$, which is known at the BS.

Denote by $K_0$ the set of the devices uploading their measurements. To be precise, let

$$K_0 = \{k_1, \ldots, k_N\}, \quad k_n \in K,$$

where $k_n$ represents the index of the $n$th node uploading local measurements. The subvector of $x$ associated with $K_0$ is denoted by $w$ and let $\Psi = [b_{k_1} \cdots b_{k_N}]^T$. Then, it can be shown that

$$w = \Psi s \in \mathbb{R}^{N \times 1},$$

which can be seen as a random sensing or sampling of $x$. It is known that if $N \geq CS \log \left(\frac{M}{K}\right)$, where $C$ is a constant, $s$ can be recovered from $w$ under certain conditions of $\Psi$ in (7) [9] [10] [17]. Once $s$ is available, $x$ can be obtained from (5). In other words, without collecting all the measurements from $K$ nodes, it is possible to estimate $x$ from $N$ nodes’ measurements.

Fig. 3 shows reconstruction errors at the BS with DAS and repeated random sensing (RRS) (i.e., random polling) after 4 rounds together with the target signal, $x$, when $K = 64$, $M = 25$, $S = 3$, and $N = 5$. As shown in Fig. 3, DAS can provide a good estimate of $x$ at the BS with $4N = 20$ measurements, while RRS cannot provide a reasonable estimate.

Let $v(t)$ and $s(t)$ be the estimates of $x$ and $s$ at round $t$ (with data set $\mathcal{N}(t)$), respectively. Note that once $s(t)$ is obtained, we have

$$v(t) = [b_1 \cdots b_K]^T s(t).$$

In DAS, for given $K(t)$, the node with the most significant measurement value (in terms of the amplitude) in round $t+1$ can be chosen as follows:

$$k(t+1) = \arg \max_{k \in K(t)} |v(t)|^2_k = \arg \max_{k \in K(t)} |b_k^T s(t)|^2.$$

Note that $x_k$ and $x_i$ are highly correlated with each other if the correlation of $b_k$ and $b_i$ is high. This can be taken into account and the resulting node section criterion for DAS becomes

$$k(t+1) = \arg \max_{k \in K(t)} \min_{i \in K(t)} \frac{|b_k^T s(t)|^2}{|b_i^T b_i|^2}.$$

While only one node is chosen per round in (9), we can choose up to $L$ nodes if there are $L$ parallel multiple access channels. In Fig. 4, the MSE of the estimate of $x$ is shown as a function of round $t$ with $K = 300$, $S = 10$, and $L = 10$ parallel multiple access channels per round. Clearly, for a reasonable estimate of $x$, it is shown that DAS needs a smaller number of rounds than RRS. It is noteworthy that the BS needs to send request signals to the selected nodes in each round through downlink channels and there might be errors in downlink transmissions (which results in uploading from unselected nodes). As shown in Fig. 4, the performance degradation due to downlink errors is not significant.

V. DISTRIBUTED DAS WITH RANDOM ACCESS

In centralized DAS, the BS is to collect measurements from devices through iterations or multiple rounds. However, unlike centralized DAS, each device decides whether or not it uploads its measurement by itself in distributed DAS. Thus, distributed DAS can be used if the BS does not have any
be the probability that node $k$ can successfully transmit its measurement without collision. If there are multiple users that choose the same channel, there is collision and no user succeeds to upload its measurement. Then, we have

$$q_k = p_k \prod_{l \neq k} \left(1 - \frac{p_l}{L}\right).$$

In order to decide $p_k$ for given local measurement $x_k$ at node $k$, consider $y$ after iteration $t$ as follows:

$$y = y_1 + y_2,$$

where $y_1 = G_1x_1$ with $x_1$ corresponding to $\mathcal{X}(t)$ and $y_2 = G_2x_2$ with $x_2$ corresponding to $\mathcal{X}^c(t)$. While $y_1$ is known at the BS, we now consider an approximation of $y_2$ with $u$ from new uploaded measurements, which is given by

$$u = \sum_{k \in \mathcal{K}^c(t)} w_k \delta_k,$$

where $w_k = g_k x_k$, and $\delta_k \in \{0, 1\}$ becomes 1 if the BS receives $x_k$ successfully and 0 otherwise. The average number of successful uploading with $L$ channels become

$$\sum_k q_k = \sum_p p_k \prod_{l \neq k} \left(1 - \frac{p_l}{L}\right) \leq \sum_p p_k \exp\left(-\sum_{l \neq k} \frac{p_l}{L}\right) \approx \sum_p p_k e^{-\frac{p}{L}} = Pe^{-\frac{P}{L}},$$

where $P = \sum_k p_k$. Thus, $P$ is to be $L$ for maximizing the average number of successful uploading, i.e.,

$$\sum_k q_k \leq Le^{-1}.$$  

Using the triangle inequality, the expectation of the error norm is bounded as

$$\mathbb{E}[||y_2 - u||] = \mathbb{E}[||\sum_{k \in \mathcal{K}^c(t)} w_k (1 - \delta_k)||] \leq \sum_{k \in \mathcal{K}^c(t)} ||w_k|| \mathbb{E}[|1 - \delta_k|] = \sum_{k \in \mathcal{K}^c(t)} ||w_k|| (1 - q_k) \leq \sum_{k \in \mathcal{K}^c(t)} ||w_k|| e^{-q_k},$$

where the last inequality is due to $1 - x < e^{-x}$, $x \geq 0$. From (16), we can have the following optimization problem:

$$\min_{\{q_k\}} \sum_{k \in \mathcal{K}^c(t)} ||w_k|| e^{-q_k}$$

subject to $\sum_{k \in \mathcal{K}^c(t)} q_k \leq Le^{-1},$ $q_k \in [0, e^{-1}]$.  

When $\sum_k p_k = L$ to maximize the average number of uploading, it can be shown that

$$q_k = p_k \prod_{l \neq k} \left(1 - \frac{p_l}{L}\right) \approx p_k e^{-\frac{\sum_{l \neq k} p_l}{L}} \approx p_k e^{-1},$$

Then, the optimization problem can be modified as

$$\min_{\{p_k\}} \sum_{k \in \mathcal{K}^c(t)} ||w_k|| \exp(-e^{-1}p_k)$$

Suppose that there are $L$ parallel multiple access channels for multichannel ALOHA. Each user can randomly choose one of $L$ channels if it decides to upload its measurement. Let $p_k$ denote the access probability of node $k$. Let $q_k$ knowledge of measurements (e.g., second order statistics of $x$ in Gaussian DAS [7] or measurement vectors in DAS for sparse signals [6]). In distributed DAS, since the BS does not decide uploading orders, random access can be used with the access probability of each node that is decided by its local measurement. In this section, we show how random access can be integrated with DAS using multichannel ALOHA [18] [19].

Suppose that the BS wants to find $y$ that is given by

$$y = G(x),$$

where $G(\cdot)$ is a certain function. For example, a linear combination can be considered as follows:

$$y = \sum_{k=1}^{K} g_k x_k = Gx.$$

Let

$$||g_{q(1)} x_{q(1)}||^2 \geq \ldots \geq ||g_{q(K)} x_{q(K)}||^2,$$

where $q(k)$ represents the index of the node that has the $k$th largest norm. Then, the uploading order can be decided according to $\{q(k)\}$. Assume that the BS knows $\{g_k\}$, but not $\{x_k\}$ that are available at nodes. If the BS knows the second order statistics of the $x_k$'s, $||g_k||^2 \mathbb{E}[||x_k||^2]$ can replace $||g_k||^2 ||x_k||^2$ and the uploading order for DAS can be decided by the BS. However, the BS may not have any information of the $x_k$'s, and, furthermore, there might be some nodes without any meaningful measurements (i.e., $||x_k|| \leq \epsilon$). Thus, it would be desirable to decide the uploading order by the nodes (i.e., distributed DAS is desirable). In distributed DAS, since the BS does not decide which devices are to transmit, random access can be employed for uploading, where the access probability can be decided by each device (based on its local measurement).
subject to $\sum_{k \in K^x(i)} p_k \leq L, \ p_k \in [0, 1].$ \hfill (19)

It can be shown that

$$p_k = \left[ e \ln \|w_k\| - \psi \right]_0^{1}, \hfill (20)$$

where $\psi$ is a Lagrange multiplier, which can be decided as

$$\psi_{t+1} = \psi_t + \mu(P_k - L), \hfill (21)$$

based on dual ascent \cite{20}. Here, $P_k$ is the number of the nodes that transmit measurements (regardless of collision), which is available at the BS after each iteration. The BS sends $\psi$ through feedback channels to the devices. Devices find their access probability as in (20) and upload their measurements accordingly.

For simulations, consider Bernoulli-Gaussian for $x_k$:

$$x_k = \begin{cases} \mathcal{N}(0, 1), & \text{w.p. } p_s, \\ \mathcal{N}(0, 1), & \text{w.p. } 1 - p_s, \end{cases}$$

where $p_s$ denotes the probability of significant measurement (i.e., $x_k$ has a large amplitude with probability $p_s$). In this case, with probability $1 - p_s$, a node has a negligible measurement. In addition, let

$$|G|_{m,k} \sim \mathcal{N}(0, 1).$$

Two different distributed DAS approaches are considered as follows:

- Random Access 1 with measurement independent access probability, $p_k = \frac{1}{|\mathcal{G}|}$
- Random Access 2 with measurement dependent access probability, $p_k = [e \ln \|w_k\| - \psi]_0^{1}$

Certainly, we expect that Random Access 2 can perform better than Random Access 1 if the access probability is optimized.

In Fig. 5, the estimation errors, $\|Y - Y(t)\|$, are shown as functions of iterations, $t$, with $K = 400, L = 10, p_s = 0.25$, and $\mu = 0.1$. As expected, Random Access 2, which has optimized access probabilities according to their local measurements, outperforms Random Access 1.

In distributed DAS, multichannel random access was considered, since the BS does not decide uploading order. By forming an optimization problem, it was able to decide the access probability at each device for efficient uploading with a limited bandwidth.

While we mainly focused on introducing DAS in terms of sensing and communication in this paper, we did not discuss its practical applications in details. We believe that DAS can be used for a number of different applications. For example, for an application that provides user’s journey time, it may require to know traffic conditions that can be estimated from real-time updating of users’ locations (or vehicle’s locations). Since it may not be possible to collect the data set of all users’ locations in real-time (due to bandwidth limitation, a large number of users, communication constraints, and so on), DAS can be employed to predict the journey time of a requested trip from a small number of selected users’ locations that affect the journey time. There would be a number of applications where DAS can efficiently provide answers to real-time queries with sensors’ or devices’ local data sets.

![Figure 5](image.png)

**Figure 5.** Estimation Errors as functions of iterations with $K = 400, L = 10, p_s = 0.25$, and $\mu = 0.1$.

In \cite{8}, we apply distributed DAS to federated learning where devices upload their local updates through parallel multiple access channels. It is shown that Random Access 2 results in fast federated learning.

**VI. CONCLUDING REMARKS**

In this paper, we presented the key idea of DAS where sensing and communication meet for efficient uploading from a large number of devices for certain real-time IoT applications including federated learning. We also discussed two different types of DAS: centralized and distributed DAS. In centralized DAS, it was assumed that a BS has certain knowledge of local measurements (e.g., second order statistics of devices’ measurements) to decide uploading order. On the other hand, in distributed DAS, the BS does not need to have any knowledge of local measurements.

In distributed DAS, multichannel random access was considered, since the BS does not decide uploading order. By forming an optimization problem, it was able to decide the access probability at each device for efficient uploading with a limited bandwidth.

While we mainly focused on introducing DAS in terms of sensing and communication in this paper, we did not discuss its practical applications in details. We believe that DAS can be used for a number of different applications. For example, for an application that provides user’s journey time, it may require to know traffic conditions that can be estimated from real-time updating of users’ locations (or vehicle’s locations). Since it may not be possible to collect the data set of all users’ locations in real-time (due to bandwidth limitation, a large number of users, communication constraints, and so on), DAS can be employed to predict the journey time of a requested trip from a small number of selected users’ locations that affect the journey time. There would be a number of applications where DAS can efficiently provide answers to real-time queries with sensors’ or devices’ local data sets.

**REFERENCES**

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, “Internet of things (IoT): A vision, architectural elements, and future directions,” Future Gener. Comput. Syst., vol. 29, pp. 1645–1660, Sept. 2013.

[2] J. Kim, J. Yun, S. Choi, D. N. Seed, G. Lu, M. Bauer, A. Al-Huzami, K. Campowsky, and J. Song, “Standard-based IoT platforms interworking: implementation, experiences, and lessons learned,” IEEE Communications Magazine, vol. 54, pp. 48–54, July 2016.

[3] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, “Internet of Things: A survey on enabling technologies, protocols, and applications,” IEEE Communications Surveys Tutorials, vol. 17, pp. 2347–2376, Fourthquarter 2015.

[4] N. Mangalvedhe, R. Ratasuk, and A. Ghosh, “NB-IoT deployment study for low power wide area cellular IoT,” in 2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), pp. 1–6, Sep. 2016.

[5] 3GPP TS 36.321 V13.2.0, Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification, June 2016.

[6] J. Choi, “A cross-layer approach to data-aided sensing using compressive random access,” IEEE Internet of Things J., vol. 6, pp. 7093–7102, Aug 2019.

[7] J. Choi, “Gaussian data-aided sensing with multichannel random access and model selection,” IEEE Internet of Things Journal, pp. 1–1, 2019.

[8] J. Choi and S. R. Pokhrel, “Federated learning with multichannel ALOHA,” IEEE Wireless Communications Letters, pp. 1–1, 2019.

[9] D. Donoho, “Compressed sensing,” IEEE Trans. Information Theory, vol. 52, pp. 1289–1306, April 2006.
[10] E. Candes, J. Romberg, and T. Tao, “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information,” IEEE Trans. Information Theory, vol. 52, pp. 489–509, Feb 2006.

[11] J. Konecný, H. B. McMahan, D. Ramage, and P. Richtárik, “Federated optimization: Distributed machine learning for on-device intelligence,” ArXiv, vol. abs/1610.02527, 2016.

[12] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” ACM Trans. Intell. Syst. Technol., vol. 10, pp. 12:1–12:19, Jan. 2019.

[13] J. Park, S. Samarakoon, M. Bennis, and M. Debbah, “Wireless network intelligence at the edge,” Proceedings of the IEEE, vol. 107, pp. 2204–2239, Nov 2019.

[14] A. Fragkiadakis, I. Askoxylakis, and E. Tragos, “Joint compressed-sensing and matrix-completion for efficient data collection in WSNs,” in 2013 IEEE 18th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), pp. 84–88, Sept 2013.

[15] C. Karakus, A. C. Gurbuz, and B. Tavli, “Analysis of energy efficiency of compressive sensing in wireless sensor networks,” IEEE Sensors Journal, vol. 13, pp. 1999–2008, May 2013.

[16] F. Wu, K. Yang, R. Duan, and T. Tian, “Compressive sampling and reconstruction of acoustic signal in underwater wireless sensor networks,” IEEE Sensors Journal, vol. 18, pp. 5876–5884, July 2018.

[17] Y. C. Eldar and G. Kutyniok, Compressed Sensing: Theory and Applications. Cambridge University Press, 2012.

[18] D. Shen and V. O. K. Li, “Performance analysis for a stabilized multi-channel slotted ALOHA algorithm,” in Proc. IEEE PIMRC, vol. 1, pp. 249–253 Vol.1, Sept 2003.

[19] C. H. Chang and R. Y. Chang, “Design and analysis of multichannel slotted ALOHA for machine-to-machine communication,” in Proc. IEEE GLOBECOM, pp. 1–6, Dec 2015.

[20] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” Found. Trends Mach. Learn., vol. 3, pp. 1–122, Jan. 2011.