

# Node Activation for SI-based xG Localization: 3GPP Case Studies using xG-Loc Dataset

Carlos A. Gómez-Vega\*, Gianluca Torsoli\*, Moe Z. Win<sup>†</sup> and Andrea Conti\*

\*Department of Engineering and CNIT, University of Ferrara, Via Saragat 1, 44122 Ferrara, Italy  
(e-mail: cgomez@ieee.org, gianluca.torsoli@unife.it, a.conti@ieee.org)

<sup>†</sup>Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 USA  
(e-mail: moewin@mit.edu)

**Abstract**—Location awareness is crucial for emerging services and enhanced resource orchestration in next generation (xG) wireless networks. To provide efficient high-accuracy localization, xG networks require both algorithms for position inference and strategies for resource optimization. While the exploitation of soft information (SI) provides significant gains in localization performance, node activation strategies benefit resource utilization by selecting an adequate set of nodes to perform measurements. This paper develops a data-driven node activation strategy for efficient SI-based localization in xG networks. First, we formulate the node activation problem considering an SI-based position estimator. Then, we propose a data-driven node activation strategy for determining an adequate set of active nodes given only a position estimate. To validate the proposed strategy, we employ xG-Loc, a dataset for location-aware xG networks fully compliant with 3rd Generation Partnership Project (3GPP) specifications. Case studies in 3GPP scenarios show the benefits of the proposed node activation strategy.

**Index Terms**—Localization, network operation, node activation, data-driven, machine learning.

## I. INTRODUCTION

Location awareness is crucial for a myriad of emerging services in next generation (xG) wireless networks [1]–[4], including autonomy [5], assets tracking [6], and Internet-of-Things [7]. The 3rd Generation Partnership Project (3GPP) providing technical specifications for cellular networks has defined use cases and performance requirements for seven positioning service levels in terms of accuracy, availability, and latency [8], [9]. Location-aware xG networks must fulfill such performance requirements with an efficient resource utilization. Nonetheless, providing efficient high-accuracy localization is difficult due to the underlying tradeoffs between attainable performance and resource utilization.

The performance of location-aware networks depends on the propagation conditions, wireless resources, and nodes deployment [10]. On the one hand, location-aware networks require algorithms for accurate position inference in complex wireless environments [11]–[13]. In particular, the exploitation of soft information (SI) has been shown to outperform existing localization algorithms in xG networks due to a thorough statistical characterization of the relationships between measurements, positional features, and contextual data [14], [15]. On the other hand, location-aware networks require strategies for efficient resource utilization [16], [17]. Such strategies allow attaining adequate tradeoffs between localization accuracy and resource

utilization, e.g., via the allocation of wireless resources and the coordination of nodes transmissions.

Efficient position inference can be obtained by performing measurements with a suitable subset of nodes [18]–[22]. In location-aware networks, node activation strategies focus on selecting an adequate set of active nodes for accurate localization [23]–[25]. However, determining the best set of active nodes for accurate localization requires channel state information and incurs significant overhead compromising latency requirements. While there are extensive studies on scheduling and handover algorithms for communications, they are inefficient or even infeasible for localization due to the different performance metrics in their design. This calls for node activation strategies that provide near-optimal decisions in terms of localization performance while employing limited information. In this regard, data-driven approaches are encouraging since they enable effective decision making in complex environments with partial information by learning from training examples [26]–[28].

The goal of this paper is twofold. First, it aims at developing a node activation strategy for efficient SI-based localization in xG networks. Second, it aims at demonstrating the capabilities of xG-Loc [29], [30], a set of 3GPP-compliant datasets for the development and evaluation of location-aware xG networks. The key idea is to learn probabilistic node activation configurations considering an SI-based position estimator.

This paper develops a data-driven node activation strategy for efficient SI-based xG localization. To validate the proposed strategy, we present case studies in 3GPP scenarios using the xG-Loc dataset. The key contributions of this paper are:

- development of a data-driven node activation strategy for efficient SI-based xG localization;
- quantification of the benefits provided by the proposed strategy in 3GPP scenarios; and
- demonstration of xG-Loc capabilities to support the development of algorithms for location-aware xG networks.

The remaining sections are organized as follows: Section II provides an overview of 3GPP-compliant localization and xG-Loc dataset. Section III formulates the node activation problem for SI-based localization. Section IV describes the proposed node activation strategy. Section V presents case studies in 3GPP scenarios. Finally, Section VI gives our conclusion.

*Notations:* Random variables are displayed in sans serif, upright fonts; their realizations in serif, italic fonts. Vectors are denoted by bold lowercase letters. For example, a random variable and its realization are denoted by  $\mathbf{x}$  and  $x$ , respectively; a random vector and its realization are denoted by  $\mathbf{x}$  and  $\mathbf{x}$ , respectively. Sets are denoted by calligraphic font. For example, a set is denoted by  $\mathcal{X}$ . The  $m$ -dimensional vector of ones is denoted by  $\mathbf{1}_m$ ; the subscript is removed when the dimension of the vector is clear from the context. The function  $f_{\mathbf{x}}(\mathbf{x}; \boldsymbol{\theta})$  indicates the probability distribution functions (PDFs) of a continuous random vector  $\mathbf{x}$  parametrized by  $\boldsymbol{\theta}$ .

## II. PRELIMINARIES

This section provides a brief overview of 3GPP radio access technology (RAT)-dependent localization and xG-Loc dataset.

### A. 3GPP RAT-dependent localization

In xG networks, the positions of user equipments (UEs) can be estimated using measurements obtained by exchanging signals with base stations (BSs) [31] and possibly neighbor UEs [32]. The 3GPP has defined two dedicated reference signals (RSs) for RAT-dependent localization, namely the positioning reference signal (PRS) in downlink (DL) and the sounding reference signal (SRS) in uplink (UL) [33]. Both RSs can be transmitted in frequency range 1 (FR1) with carrier frequency between 410 MHz and 7.125 GHz or frequency range 2 (FR2) with carrier frequency between 24.25 GHz and 52.6 GHz. The processing of the received RSs allows to extract relevant power-, time-, and angle-based measurements for localization, including time-of-arrival (TOA), time-difference-of-arrival (TDOA), round-trip time (RTT), and angle-of-departure (AOD) [31]. In addition, the detection and identification of line-of-sight (LOS)/non-line-of-sight (NLOS) condition can be exploited for xG localization [32].

### B. xG-Loc Dataset

xG-Loc is an open-source collection of 3GPP-compliant datasets for the development and evaluation of localization algorithms and location-based services [29], [30]. xG-Loc includes received RSs, time- and angle-based measurements, and analytics for different network and signal settings in 3GPP indoor and outdoor scenarios with carrier frequencies in both FR1 and FR2. The xG-Loc dataset also provides summary data in JavaScript object notation (JSON) files that can be used without requiring further processing of the received RSs [29]. The summary data in xG-Loc JSON files include:

- ground truth UE position;
- ground truth LOS/NLOS indicator;
- UE position estimate obtained via SI-based localization considering RTT measurements [14];
- range estimates from UL-TOA and DL-TOA measurements obtained by processing the received SRS and PRS, respectively;
- wireless channel quality indicator obtained via blockage intelligence (BI) [15]; and
- AOD estimates for the indoor scenarios in FR2.

## III. PROBLEM FORMULATION

This section formulates the node activation problem considering an SI-based position estimator.

### A. SI-based localization

Consider a non-cooperative location-aware xG network composed of a single UE with unknown position and  $N_b$  BSs with known positions.<sup>1</sup> The index set of BSs is denoted by  $\mathcal{N}_b = \{1, 2, \dots, N_b\}$ . The positions of the UE and BSs are denoted by  $\mathbf{p}$  and  $\mathbf{p}_j$  for  $j \in \mathcal{N}_b$ , respectively. The goal is to estimate the UE position by leveraging a collection of measurements  $\{\mathbf{y}_j\}_{j \in \mathcal{N}_b}$ . Specifically, each measurement  $\mathbf{y}_j$  obtained between the UE and the BS  $j$  is related to a positional feature  $\boldsymbol{\theta}_j(\mathbf{p})$ . Such measurements can include received waveform samples, power-, time-, or angle-based metrics, or any combination of them.

SI-based localization algorithms exploit a statistical characterization of the relationships between measurements, positional features, and contextual data [12], [14]. The SI of a positional feature  $\boldsymbol{\theta}$  encapsulated in a measurement  $\mathbf{y}$  is denoted by  $\mathcal{L}_{\mathbf{y}}(\boldsymbol{\theta})$ . In a non-Bayesian setting, we have  $\mathcal{L}_{\mathbf{y}}(\boldsymbol{\theta}) \propto f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\theta})$ . In particular, SI-based localization is divided in two phases, namely offline training and online operation.

In the training phase, a generative model for  $\mathcal{L}_{\mathbf{y}}(\boldsymbol{\theta})$  is learned from offline measurements [12], [14], [15]. The generative model encapsulates the position information contained in measurements based on their joint distribution function, which can be approximated via density estimation. An effective approach to learn SI relies on fitting a Gaussian mixture model (GMM) using the expectation-maximization algorithm [34].

In the operation phase, the SI is determined online based on the generative models learned offline and the new measurements collected. Such SI is then employed to estimate the UE position. Let  $\mathcal{N}_s \subseteq \mathcal{N}_b$  denote the index set of active BSs providing measurements. The UE position can be obtained via maximum likelihood (ML) estimation as

$$\hat{\mathbf{p}} = \arg \max_{\tilde{\mathbf{p}}} \prod_{j \in \mathcal{N}_s} \mathcal{L}_{\mathbf{y}_j}(\boldsymbol{\theta}_j(\tilde{\mathbf{p}})). \quad (1)$$

### B. Node activation problem

Node activation strategies aim to determine the set of active BSs  $\mathcal{N}_s$  that minimizes the position error. In particular, determining the best set of active BS requires information about the quality of the measurements obtained. To formulate the node activation problem, consider that all the available BSs provide measurements for localization. Let  $\mathbf{u} = [u_1, u_2, \dots, u_{N_b}]^T$  denote the node activation vector (NAV) with  $u_j \in \{0, 1\}$  representing whether the measurement from BS  $j$  is selected ( $u_j = 1$ ) or not ( $u_j = 0$ ). The UE position estimate can be expressed as a function of the NAV as

$$\hat{\mathbf{p}}(\mathbf{u}) = \arg \max_{\tilde{\mathbf{p}}} \sum_{j \in \mathcal{N}_b} u_j \mathcal{L}_{\mathbf{y}_j}(\boldsymbol{\theta}_j(\tilde{\mathbf{p}})) \quad (2)$$

<sup>1</sup>In fifth generation (5G) networks, the BSs are referred to as gNodeBs (gNBs).

where  $\ell_{y_j}(\theta_j(\tilde{\mathbf{p}})) = \ln\{\mathcal{L}_{y_j}(\theta_j(\tilde{\mathbf{p}}))\}$  denotes the logarithmic SI. The terms in the summation are the SI contributions provided by the measurements collected with respect to each BS, which are selected according to the activation variables  $u_j$ . The position error as a function of the NAV is given by

$$e(\mathbf{u}; \mathbf{p}) = \|\hat{\mathbf{p}}(\mathbf{u}) - \mathbf{p}\|. \quad (3)$$

Considering the position error (3) as performance metric, the node activation problem for a UE at a given position  $\mathbf{p}$  can be formulated as

$$\mathcal{P}_{\mathbf{p}} : \underset{\mathbf{u}}{\text{minimize}} \quad e(\mathbf{u}; \mathbf{p}) \quad (4a)$$

$$\text{subject to} \quad \mathbf{1}^T \mathbf{u} = N_s \quad (4b)$$

$$u_j \in \{0, 1\}, \quad j \in \mathcal{N}_b \quad (4c)$$

where  $N_s$  is a parameter indicating the number of BSs to be activated, i.e.,  $N_s = |\mathcal{N}_s|$ . In  $\mathcal{P}_{\mathbf{p}}$ , (4b) describes the constraint on the total number of active BSs, i.e.,  $\sum_{i=1}^{N_b} u_j = N_s$ , and (4c) represents the Boolean constraints for each  $u_j$ . The optimal solution to  $\mathcal{P}_{\mathbf{p}}$  is denoted by  $\mathbf{u}^* = [u_1^*, u_2^*, \dots, u_{N_b}^*]^T$ . Such a solution provides the optimal set of active BSs denoted by  $\mathcal{N}_s^* = \{j : u_j^* = 1\}$ . Note that problem  $\mathcal{P}_{\mathbf{p}}$  depends on the generative models used for the SI-based position estimator. Therefore, the optimal solution to  $\mathcal{P}_{\mathbf{p}}$  is tailored to the employed SI-based localization algorithm.

*Remark 1:* Solving  $\mathcal{P}_{\mathbf{p}}$  in (4) is challenging since the cost function is non-convex in general. In particular, solving this combinatorial problem via exhaustive search requires evaluating  $\binom{N_b}{N_s}$  possible solutions. Note also that the problem requires prior knowledge about the UE position as well as measurements from all the BSs in the network. Therefore, efficient localization calls for developing efficient node activation strategies that solve  $\mathcal{P}_{\mathbf{p}}$  approximately by employing limited information during online operation.

#### IV. DATA-DRIVEN NODE ACTIVATION STRATEGY

This section describes the proposed node activation strategy. Specifically, we consider a data-driven approach to learn a probabilistic node activation decision rule from training data obtained by solving  $\mathcal{P}_{\mathbf{p}}$ .

##### A. Offline training

In the offline phase, we aim at determining a decision rule to provide near-optimal node activation configurations given a position estimate  $\hat{\mathbf{p}}$ . Consider a classification problem consisting of  $N_b$  classes, each of which representing the activation of a single BS. Solving this problem allows determining a vector of activation probabilities  $\tilde{\mathbf{u}} = [\tilde{u}_1, \tilde{u}_2, \dots, \tilde{u}_{N_b}]^T$  for soft node activation such that  $\tilde{u}_j \geq 0$  and  $\sum_{j=1}^{N_b} \tilde{u}_j = 1$ . Specifically, such probabilities can be computed via a parametric mapping obtained offline in a supervised learning setting with training data collected by solving instances of  $\mathcal{P}_{\mathbf{p}}$ .

For a given instance of  $\mathcal{P}_{\mathbf{p}}$ , we break the optimal NAV into  $N_s$  vectors indicating the activation of a single BS. Considering the one-hot encoding scheme [34], we have an  $N_b$ -dimensional vector whose elements are all zero except the

element  $j$  corresponding to the activation of the BS  $j$ , which takes the value 1. Thereby, each instance of the problem  $\mathcal{P}_{\mathbf{p}}$  with optimal solution  $\mathbf{u}^*$  provides  $N_s$  training examples of the form  $(\hat{\mathbf{p}}, \mathbf{u}_j)$  for  $j$  such that  $u_j^* = 1$  where  $\mathbf{u}_j$  denotes the one-hot encoded vector for the BS  $j$ .

Consider the state and decision spaces denoted by  $\mathcal{X}$  and  $\mathcal{U}$ , respectively, such that  $\hat{\mathbf{p}} \in \mathcal{X}$  and  $\tilde{\mathbf{u}} \in \mathcal{U}$ . Let  $g : \mathcal{X} \mapsto \mathcal{U}$  denote a mapping from the state space to the decision space. We consider a family of parametric mappings  $\mathcal{G}$  given by a predefined neural network architecture with parameter space  $\Psi$  [35]. Note that each  $\psi \in \Psi$  defines a different mapping  $g(\cdot; \psi) \in \mathcal{G}$ . Specifically, we consider a fully connected neural network architecture to perform the classification task.

The goal of the offline phase is to determine the parameters  $\psi^* \in \Psi$  that provide the best fit to training data. Let  $\{(\hat{\mathbf{p}}^{(m)}, \tilde{\mathbf{u}}^{(m)})\}_{m \in \mathcal{N}_{\text{train}}}$  denote the training data indexed by  $\mathcal{N}_{\text{train}}$ . Since we consider a classification problem, the output layer of the neural network employs softmax activation functions enabling the probabilistic interpretation of a categorical distribution of  $N_b$  elements, and the objective function to fit the model is the cross-entropy loss given the training data [34], [35]. By minimizing the cross-entropy loss in the training phase, the neural network is encouraged to match the labels of the training data and approximate the desired distribution. Note that the role of the neural network in the node activation strategy is to approximate a decision rule for a probabilistic classification task in the form of a parametric function.

To perform offline training more efficiently we consider the following procedures.

1) *Search space reduction:* Solving the combinatorial problem  $\mathcal{P}_{\mathbf{p}}$  via exhaustive search requires substantial computation. While the offline phase is performed only once to learn the decision rule, the reduction of its complexity is crucial for an amenable training phase. We reduce the search space of  $\mathcal{P}_{\mathbf{p}}$  by exploiting the channel quality indicators provided by BI [15]. BI provides a probabilistic indicator of LOS condition that takes into account statistical features of the received waveform, hence also encapsulating information about the channel quality. The use of BI allows us to discard BSs in a principled manner by considering the probability of being in NLOS condition with poor channel quality. Specifically, we consider the BSs with the best  $N_r$  BI values, discarding the  $N_b - N_r$  BSs with the worst channel conditions. By employing BI to reduce the search space, solving  $\mathcal{P}_{\mathbf{p}}$  requires evaluating  $\binom{N_r}{N_s}$  possible solutions instead of  $\binom{N_b}{N_s}$ .

2) *Data augmentation:* The performance of the decision rule determined offline can be affected by unpredictable position errors due to either unreliable measurements or sub-optimal node activation configurations. We consider data augmentation with factor  $N_{\text{aug}}$  to increase the amount of training instances and obtain a decision rule that is more robust to position errors. This procedure consists of generating  $N_{\text{aug}} - 1$  new training instances by considering position estimates obtained with different node activation configurations. Such estimates are paired with the optimal node activation decisions for the ground truth position, e.g.,  $(\hat{\mathbf{p}}, \tilde{\mathbf{u}})$ .

### B. Online operation

In the online phase, the node activation configuration is obtained by evaluating the parametric mapping determined offline using a position estimate  $\hat{\mathbf{p}}$ . For example, the first position estimate can be obtained by performing measurements with all the available BSs.

From the training phase, we determine the optimal parameter  $\psi^*$  to estimate activation probabilities online by evaluating the parametric function defined by the neural network architecture. Given a position estimate  $\hat{\mathbf{p}}$ , we obtain the activation probabilities online as

$$\tilde{\mathbf{u}} = g(\hat{\mathbf{p}}; \psi^*). \quad (5)$$

From (5), we can select the  $N_s$  BSs with the highest probabilistic scores. Let  $\hat{\mathbf{u}} = [\hat{u}_1, \hat{u}_2, \dots, \hat{u}_{N_b}]^T$  denote the estimated NAV. In particular, the elements of  $\hat{\mathbf{u}}$  are given by

$$\hat{u}_j = \begin{cases} 1 & \text{if } \tilde{u}_j \text{ is one of the } N_s \text{ largest elements in } \tilde{\mathbf{u}} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where ties are broken arbitrarily. Note that the online computation requires to evaluate the parametric function (a feedforward pass of a neural network) and sorting the activation probabilities to find the best  $N_s$ . Therefore, the complexity of online operation is determined by the neural network architecture, i.e., by the number and size of its hidden layers.

### V. CASE STUDIES

This section evaluates the performance of the proposed data-driven node activation strategy in 3GPP scenarios using the xG-Loc dataset [29], [30].

We consider 3GPP indoor open office (IOO) and indoor factory (InF)-sparse high (SH) scenarios [36]. The IOO scenario consists of an environment of 120 m  $\times$  50 m where  $N_b = 12$  BSs are deployed. The InF-SH scenario consists of an environment of 300 m  $\times$  150 m where  $N_b = 18$  BSs are deployed. We consider the xG-Loc configurations employing RSs transmitted with 100 MHz bandwidth at carrier frequencies of 4 GHz and 3.5 GHz for the IOO and InF-SH scenarios, respectively [29], [30]. For each configuration, xG-Loc includes a total of 100 instances of the simulation scenario with random parameters. Each instance considers 10 UEs deployed in the scenario with random position and orientation. Therefore, the dataset contains a total of  $N_d = 1000$  samples. In particular, we only leverage the summary data provided in the xG-Loc JSON files without processing the received RSs.

The training phase of the proposed strategy requires knowledge of the generative models for SI-based localization. Therefore, we learn the generative models for SI before the training phase of the proposed strategy. Specifically, we adopt a GMM with 8 components and employ a 10-fold cross-validation technique to obtain the generative models [34]. Once the generative models are obtained, we collect training instances for the data-driven node activation strategy as in the following.

First, we determine the optimal configuration of  $N_s$  active BSs by solving  $\mathcal{P}_p$  (4) via exhaustive search. In particular, we

TABLE I  
PARAMETERS OF POSITION ERROR MODELS FOR DATA AUGMENTATION

| $N_s$ | $\mathbf{e}$   | IOO     |          | InF-SH  |          |
|-------|----------------|---------|----------|---------|----------|
|       |                | $\mu$   | $\sigma$ | $\mu$   | $\sigma$ |
| 3     | $\mathbf{e}_x$ | -0.1216 | 1.7888   | -0.1324 | 4.1769   |
|       | $\mathbf{e}_y$ | -0.3795 | 4.8635   | -0.6136 | 8.6975   |
| 4     | $\mathbf{e}_x$ | -0.0821 | 1.1766   | -0.0435 | 0.9453   |
|       | $\mathbf{e}_y$ | 0.0429  | 3.0816   | -0.1279 | 2.7850   |
| 5     | $\mathbf{e}_x$ | -0.8891 | 0.8003   | -0.0061 | 0.3588   |
|       | $\mathbf{e}_y$ | 0.0383  | 2.1778   | -0.0126 | 0.4768   |
| 6     | $\mathbf{e}_x$ | 0.0142  | 0.5241   | -0.0080 | 0.3291   |
|       | $\mathbf{e}_y$ | 0.0461  | 1.8223   | 0.0012  | 0.4279   |

consider  $N_s = 3, 4, 5$ , and 6 BSs to evaluate the performance of the proposed strategy. We perform a search space reduction exploiting the BI probabilistic indicator considering  $N_r = 8$  BSs. For  $N_s = 6$  in IOO scenario, solving  $\mathcal{P}_p$  with the proposed search space reduction requires the evaluation of  $\binom{8}{6}$  candidate solutions corresponding to 3% of the  $\binom{12}{6}$  possible solutions. For  $N_s = 6$  in InF-SH scenario, solving  $\mathcal{P}_p$  with the proposed search space reduction requires the evaluation of  $\binom{8}{6}$  candidate solutions corresponding to 0.1% of the  $\binom{18}{6}$  possible solutions. Hence, the search space reduction via BI enables amenable offline training.

Next, we perform data augmentation with  $N_{\text{aug}} = 5$  to generate new training instances. To this end, we fit models of the horizontal localization error on the  $x$  and  $y$  coordinates. Such models enable generating new training instances with consistent localization errors considering also the node activation strategy. Let  $\mathbf{e}_x$  and  $\mathbf{e}_y$  be the localization error on the  $x$  and  $y$  coordinates, respectively. We consider that  $\mathbf{e}_x$  and  $\mathbf{e}_y$  are Gaussian random variables and determine their parameters (i.e., mean  $\mu$  and standard deviation  $\sigma$ ) via ML estimation with training data consisting of position estimates obtained with different NAVs for each value of  $N_s$  considered. Such NAVs include the optimal solution to  $\mathcal{P}_p$  obtained via exhaustive search with search space reduction as well as sub-optimal configurations. Table I shows the parameters of the distributions obtained for the first fold of the cross-validation procedure for the different values of  $N_s$  considered. Note that the standard deviations for the models in InF-SH scenarios with  $N_s = 4, 5$ , and 6 are smaller than those in IOO scenarios, indicating less variability in the position error. This can be attributed to the larger number of gNBs available in the InF-SH scenario, as well as to a lower NLOS probability [36].

To complete the offline phase, we train fully connected neural network architectures with the augmented training data. We consider neural networks composed of 7 hidden layers with 32, 64, 128, 256, 512, 64, and 32 neurons, respectively. The input and output layers have sizes of 2 and  $N_b$ , respectively. The activation functions of the hidden layers are rectified linear units, while those of the output layer are softmax functions. The neural network architectures are trained using the Adam algorithm with 30 epochs and a batch size of 128 [35].

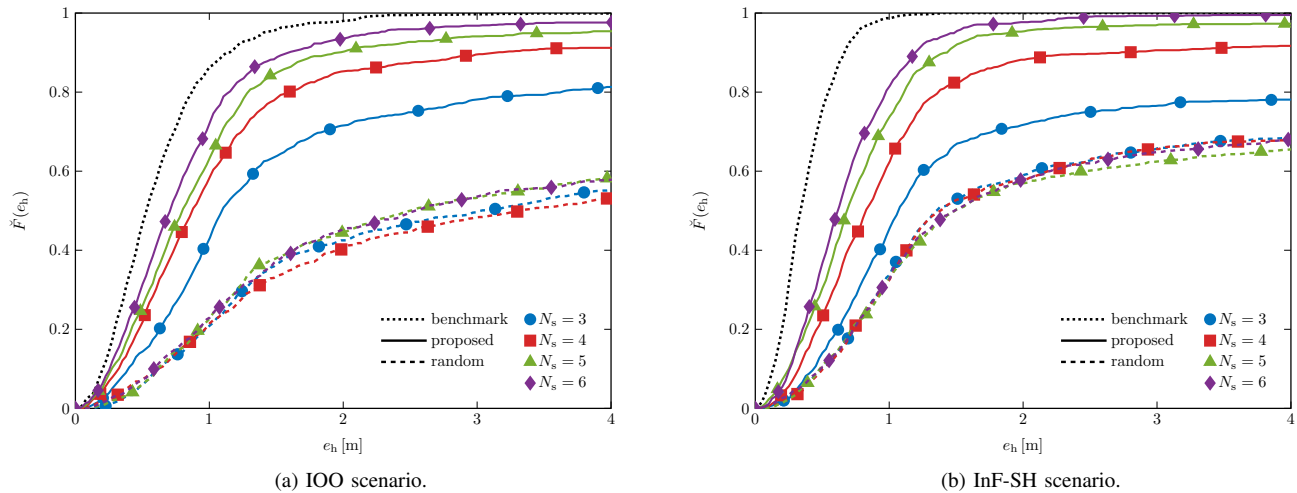


Fig. 1. Performance of node activation strategies with different numbers of active BSs in the considered 3GPP scenarios.

The performance of the proposed node activation strategy is evaluated on test data and compared with random node activation. We also consider the localization accuracy obtained by employing all the BSs in the scenario as benchmark. Note that this benchmark corresponds with the methodology in 3GPP technical reports [37]. The localization performance is evaluated in terms of the empirical cumulative distribution function (ECDF) of the horizontal localization error  $e_h$ . We denote the ECDF of  $e_h$  as  $\check{F}(e_h)$ .

Fig. 1a shows the performance of the proposed data-driven node activation strategy in the IOO scenario for different values of  $N_s$ . Note that the proposed strategy provides significant performance improvements compared to random node activation. While the random activation strategy selects BSs in both LOS and NLOS conditions randomly, the proposed strategy selects BSs with favorable channel conditions by relying on a probabilistic characterization. In particular, increasing  $N_s$  provides a performance improvement in the proposed strategy. At the 90th percentile, the position error for the baseline performance and the proposed strategy with  $N_s = 4$  and  $N_s = 6$  are 1.13 m, 3.14 m, and 1.60 m, respectively. This indicates that reducing the number of BSs, i.e., the amount of network resources employed, by 67% and 50% implies performance losses of 2.01 m and 0.47 m for  $N_s = 4$  and  $N_s = 6$  active BSs, respectively. These results show the underlying tradeoff between attainable performance and resource utilization.

Fig. 1b shows the performance of the proposed data-driven node activation strategy for in the InF-SH scenario for different values of  $N_s$ . Similarly to the IOO scenario, the proposed strategy provides significant performance improvements compared to random activation in the InF-SH scenario. Furthermore, it can be noticed that the proposed strategy provides better performance in the InF-SH scenario compared to the IOO scenario. At the 90th percentile, the position error for the baseline performance and the proposed strategy with  $N_s = 4$  and  $N_s = 6$  are 0.67 m, 2.73 m, and 1.20 m, respectively.

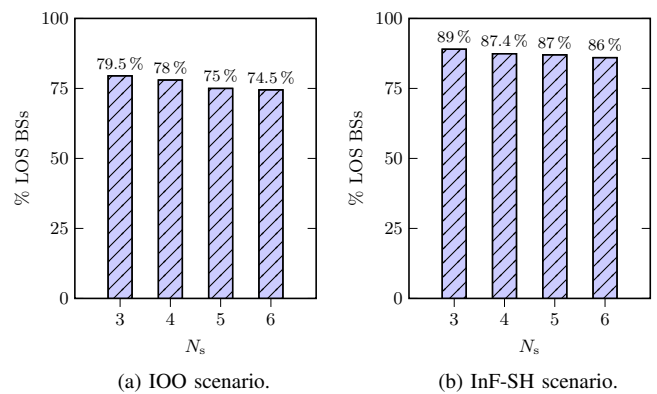


Fig. 2. Percentage of active BSs in LOS condition when selected according to the proposed strategy in 3GPP scenarios.

These results indicate that reducing the number of BSs by 78% and 67% implicate performance losses of 2.06 m and 0.53 m for  $N_s = 4$  and  $N_s = 6$  active BSs, respectively. In particular, the reduction of active BSs is more significant in the InF-SH scenario due to its higher number of BSs available.

The results on the localization error show that the proposed node activation strategy can activate BSs in favorable conditions and provide adequate performance with only the position estimate as parameter. To gain more insights into these results, Fig. 2 shows the percentage of BSs in LOS condition activated by the proposed node activation strategy for the settings considered. In the IOO scenario, the percentage of BSs in LOS condition varies between 74.5% and 79.5%. In the InF-SH scenario, the percentage of BSs in LOS condition varies between 86% and 89%. These results show that the percentage of BSs in LOS condition selected does not vary significantly with  $N_s$ . Note that the InF-SH scenario shows better performance due to a lower variability in the position error (cf. Table I). This demonstrates the effectiveness of the proposed strategy in activating BSs that are in favorable channel conditions to provide adequate localization performance.

## VI. CONCLUSION

This paper presents a data-driven node activation strategy for efficient SI-based localization in xG networks. The proposed strategy takes into account the characteristics of the SI-based position estimator and employs only a position estimate in the decision rule for node activation. In particular, this approach allows to learn NAVs tailored to the generative models used for SI-based localization. Results in 3GPP scenarios using the xG-Loc dataset show that the proposed strategy outperforms random node activation due to the activation of BSs with favorable channel conditions. The proposed strategy provides satisfactory performance with limited information, hence reducing measurement overhead for decision making. The presented results reveal the potential benefits of data-driven node activation for SI-based localization and demonstrate the capabilities of xG-Loc dataset for expediting the development of location-aware xG networks.

## ACKNOWLEDGMENT

The fundamental research described in this paper was supported, in part, by the Office of Naval Research under Grant N62909-22-1-2009, by the National Science Foundation under Grant CNS-2148251, and by federal agency and industry partners in the RINGS program.

## REFERENCES

- [1] *Technical Specification Group Services and System Aspects; Study on positioning use cases; Stage 1*, 3rd Generation Partnership Project 3GPP™ TR 22.872 V16.1.0 (2018-09), Sep. 2018, Release 16.
- [2] M. Z. Win *et al.*, “Network localization and navigation via cooperation,” *IEEE Commun. Mag.*, vol. 49, no. 5, pp. 56–62, May 2011.
- [3] W. Chen *et al.*, “5G-Advanced toward 6G: Past, present, and future,” *IEEE J. Sel. Areas Commun.*, vol. 41, no. 6, pp. 1592–1619, Jun. 2023.
- [4] A. Conti *et al.*, “Location awareness in beyond 5G networks,” *IEEE Commun. Mag.*, vol. 59, no. 11, pp. 22–27, Nov. 2021.
- [5] A. Guerra, F. Guidi, D. Dardari, and P. M. Djurić, “Networks of UAVs of low complexity for time-critical localization,” *IEEE Aerosp. Electron. Syst. Mag.*, vol. 37, no. 10, pp. 22–38, Oct. 2022.
- [6] L. Barbieri, M. Brambilla, A. Trabattani, S. Mervic, and M. Nicoli, “UWB localization in a smart factory: Augmentation methods and experimental assessment,” *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–18, Mar. 2021, Art no. 2508218.
- [7] Y. Li *et al.*, “Toward location-enabled IoT (LE-IoT): IoT positioning techniques, error sources, and error mitigation,” *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4035–4062, Mar. 2021.
- [8] *Technical Specification Group Services and System Aspects; Service requirements for the 5G system; Stage 1*, 3rd Generation Partnership Project 3GPP™ TS 22.261 V18.6.0 (2022-03), Mar. 2022, Release 18.
- [9] *Technical Specification Group Services and System Aspects; Service requirements for cyber-physical control applications in vertical domains; Stage 1*, 3rd Generation Partnership Project 3GPP™ TS 22.104 V18.3.0 (2021-12), Dec. 2021, Release 18.
- [10] M. Z. Win, Y. Shen, and W. Dai, “A theoretical foundation of network localization and navigation,” *Proc. IEEE*, vol. 106, no. 7, pp. 1136–1165, Jul. 2018.
- [11] X. Wang, L. Gao, and S. Mao, “CSI phase fingerprinting for indoor localization with a deep learning approach,” *IEEE Internet Things J.*, vol. 3, no. 6, pp. 1113–1123, Dec. 2016.
- [12] A. Conti, S. Mazuelas, S. Bartoletti, W. C. Lindsey, and M. Z. Win, “Soft information for localization-of-things,” *Proc. IEEE*, vol. 107, no. 11, pp. 2240–2264, Nov. 2019.
- [13] J. L. Carrera Villacrés, Z. Zhao, T. Braun, and Z. Li, “A particle filter-based reinforcement learning approach for reliable wireless indoor positioning,” *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2457–2473, Nov. 2019.
- [14] F. Morselli, S. M. Razavi, M. Z. Win, and A. Conti, “Soft information based localization for 5G networks and beyond,” *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 9923–9938, Dec. 2023.
- [15] G. Torsoli, M. Z. Win, and A. Conti, “Blockage intelligence in complex environments for beyond 5G localization,” *IEEE J. Sel. Areas Commun.*, vol. 41, no. 6, pp. 1688–1701, Jun. 2023.
- [16] P. Yang, C. Xiang, and S. Zhang, “Distributed joint power and bandwidth allocation for multiagent cooperative localization,” *IEEE Commun. Lett.*, vol. 26, no. 11, pp. 2601–2605, Nov. 2022.
- [17] M. Z. Win, W. Dai, Y. Shen, G. Chrisikos, and H. V. Poor, “Network operation strategies for efficient localization and navigation,” *Proc. IEEE*, vol. 106, no. 7, pp. 1224–1254, Jul. 2018.
- [18] S. Joshi and S. Boyd, “Sensor selection via convex optimization,” *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 451–462, Feb. 2009.
- [19] X. Shen and P. K. Varshney, “Sensor selection based on generalized information gain for target tracking in large sensor networks,” *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 363–375, Jan. 2014.
- [20] S. Bartoletti, A. Giorgetti, M. Z. Win, and A. Conti, “Blind selection of representative observations for sensor radar networks,” *IEEE Trans. Veh. Technol.*, vol. 64, no. 4, pp. 1388–1400, Apr. 2015.
- [21] X. Shen, S. Liu, and P. K. Varshney, “Sensor selection for nonlinear systems in large sensor networks,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 50, no. 4, pp. 2664–2678, Oct. 2014.
- [22] C. A. Gómez-Vega, Z. Liu, C. A. Gutiérrez, M. Z. Win, and A. Conti, “Efficient deployment strategies for network localization with assisting nodes,” *IEEE Trans. Mobile Comput.*, vol. 23, no. 5, pp. 6272–6287, May 2024.
- [23] X. Li, N. Mitton, I. Simplot-Ryl, and D. Simplot-Ryl, “Dynamic beacon mobility scheduling for sensor localization,” *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 8, pp. 1439–1452, Aug. 2012.
- [24] S. Dwivedi, D. Zachariah, A. De Angelis, and P. Händel, “Cooperative decentralized localization using scheduled wireless transmissions,” *IEEE Commun. Lett.*, vol. 17, no. 6, pp. 1240–1243, Jun. 2013.
- [25] T. Wang, A. Conti, and M. Z. Win, “Network navigation with scheduling: Distributed algorithms,” *IEEE/ACM Trans. Netw.*, vol. 27, no. 4, pp. 1319–1329, Aug. 2019.
- [26] Q. Mao, F. Hu, and Q. Hao, “Deep learning for intelligent wireless networks: A comprehensive survey,” *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2595–2621, Fourth Quarter 2018.
- [27] A. Zappone, M. D. Renzo, and M. Debbah, “Wireless networks design in the era of deep learning: Model-based, AI-based, or both?” *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7331–7376, Oct. 2019.
- [28] N. Shlezinger, Y. C. Eldar, and S. P. Boyd, “Model-based deep learning: On the intersection of deep learning and optimization,” *IEEE Access*, vol. 10, pp. 115 384–115 398, 2022.
- [29] A. Conti, G. Torsoli, C. A. Gómez-Vega, A. Vaccari, G. Mazzini, and M. Z. Win, “3GPP-compliant datasets for xG location-aware networks,” *IEEE Open J. Veh. Technol.*, vol. 5, pp. 473–484, Apr. 2024.
- [30] A. Conti, G. Torsoli, C. A. Gómez-Vega, A. Vaccari, and M. Z. Win, “xG-Loc: 3GPP-compliant datasets for xG location-aware networks,” *IEEE Dataport*, Dec. 2023. [Online]. Available: <https://dx.doi.org/10.21227/rper-vc03>
- [31] *Technical Specification Group Radio Access Network; Study on NR Positioning Enhancements*, 3rd Generation Partnership Project 3GPP™ TR 38.857 V17.0.0 (2021-03), Mar. 2021, Release 17.
- [32] *Technical Specification Group Radio Access Network; Study on expanded and improved NR positioning*, 3rd Generation Partnership Project 3GPP™ TR 38.859 V18.0.0 (2022-12), Dec. 2022, Release 18.
- [33] *Technical Specification Group Radio Access Network; NR; Physical channels and modulation*, 3GPP™ Std. TS 38.211 V18.1.0 (2024-01), Jan. 2024, Release 18.
- [34] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
- [35] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2017.
- [36] *Technical Specification Group Radio Access Network; Study on channel model for frequencies from 0.5 to 100 GHz*, 3GPP™ Std. TR 38.901 V17.0.0, Mar. 2022, Release 17.
- [37] *Technical Specification Group Radio Access Network; Study on NR positioning support*, 3rd Generation Partnership Project 3GPP™ TR 38.855 V16.0.0, Mar. 2019, Release 16.