Byzantine-tolerant federated Gaussian process regression for streaming data

Xu Zhang
Pennsylvania State University
xxz313@psu.edu

Zhenyuan Yuan
Pennsylvania State University
zqy5086@psu.edu

Minghui Zhu
Pennsylvania State University
muz16@psu.edu

Abstract

In this paper, we consider Byzantine-tolerant federated learning for streaming data using Gaussian process regression (GPR). In particular, a cloud and a group of agents aim to collaboratively learn a latent function where some agents are subject to Byzantine attacks. We develop a Byzantine-tolerant federated GPR algorithm, which includes three modules: agent-based local GPR, cloud-based aggregated GPR and agent-based fused GPR. We derive the upper bounds on prediction error between the mean from the cloud-based aggregated GPR and the target function provided that Byzantine agents are less than one quarter of all the agents. We also characterize the lower and upper bounds of the predictive variance. Experiments on a synthetic dataset and two real-world datasets are conducted to evaluate the proposed algorithm.

The Internet of Things (IoT) produces large volumes of spatially distributed data. A typical practice is to store data in a centralized entity, e.g., a cloud, and train machine learning models there. However, it poses challenges in communication overhead and computation efficiency. In addition, data storage on the centralized entity raises privacy concerns. Federated learning provides a promising paradigm to address the two challenges [1], [2], [3], [4]. First, only a finite number of model parameters are shared and the size of these parameters is much smaller than that of raw data; second, data samples are kept by agents and local models are trained without sharing raw data. However, these agents are vulnerable to Byzantine attacks, which force them to behave arbitrarily and send any message to the cloud. The cloud cannot access each agent's local private data and training process, and then cannot verify the correctness of each agent's updates. Byzantine attacks incur degradation of learning performance.

Related work. There have been many recent works to secure federated learning against Byzantine attacks [5], [6], [7], [8], [9]. Paper [5] proposes *Krum* as the aggregation rule where the cloud selects a gradient such that the overall distance between this gradient to a fixed number of its closest gradients is minimal. Paper [6] develops an algorithm which is based on integrated stochastic quantization, geometric median based outlier detection and secure model aggregation. Paper [7] analyzes two robust distributed gradient descent algorithms based on median and trimmed mean operations, respectively. Paper [8] proposes a high-dimensional robust mean estimation algorithm at the cloud to combat the adversary. In paper [9], the cloud partitions all the received local gradients into batches and computes the mean of each batch. After that, the cloud computes the geometric median of the batch means. The above set of papers can guarantee that the estimate of the parameter aggregated by the cloud will converge to a neighborhood of the minimizer of the loss function regardless of Byzantine attacks. However, the existing works only consider deep neural networks (DNN) as the learning model and are limited to static data and off-line learning.

On-line learning is demanded to process data that arrives sequentially across time [10], [11], [12]. Gaussian process regression (GPR) is a solution candidate to solve the problem [13], [14], [15], [16]. First of all, GPR is a non-parametric statistical learning model. With proper choice of prior covariance function and mild assumptions of the target function, GPR is able to approximate any

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

continuous function [17]. Second, GPR can be implemented recursively, reducing computational complexity and memory [18]. Third, GPR is able to quantify learning uncertainties since it predicts the target function using a posterior probability distribution [19]. However, Byzantine-tolerant GPR has not been studied yet.

Contributions. This paper considers Byzantine-tolerant federated learning for streaming data using GPR where a cloud and a group of agents aim to collaboratively learn a latent function despite some agents are subject to Byzantine attacks. Major contributions are listed as follows:

1) We develop a Byzantine-tolerant federated GPR algorithm, which includes three modules: agent-based local GPR, cloud-based aggregated GPR and agent-based fused GPR. Specifically, the agent-based local GPR sends potentially compromised local predictions to the cloud, and the cloud-based aggregated GPR computes a global model by a Byzantine-tolerant PoE aggregation rule. Then the cloud broadcasts the current global model to all the agents. Agent-based fused GPR returns the predictions by fusing the models from the cloud-based GPR with those from the agent-based local GPR.
2) The predictive mean of the cloud-based aggregated GPR is theoretically guaranteed to be restricted within a neighborhood of the target function provided that Byzantine agents are less than one quarter of all the agents. The prediction error is positively proportional to the number of Byzantine agents and that of the agents with extreme values. Prediction uncertainty is quantified by lower and upper bounds, which are identical to that of the attack-free case.

Experiments on a synthetic dataset and two real-world datasets are conducted to demonstrate that the Byzantine-tolerant GPR algorithm is resilient to Byzantine attacks. To the best knowledge of authors, this is the first paper to study Byzantine-tolerant federated learning over streaming data using GPR.

Notations and notions. Throughout the paper, we use lower-case letters, e.g, a, to denote scalars, bold letters, e.g., $\boldsymbol{a}$, to denote vectors; we use upper-case letters, e.g., A, to denote matrices, calligraphic letters, e.g., $\mathcal{A}$, to denote sets, and bold calligraphic letters, e.g., $\boldsymbol{\mathcal{A}}$, to denote spaces. Denote $I_n \in \mathbb{R}^{n\times n}$ the n-by-n identity matrix. We denote by $\mathbb{R}$ the set of real numbers. The set of non-negative real numbers is denoted by $\mathbb{R}_+$, and the set of positive real numbers is denoted by $\mathbb{R}_{++}$. The cardinality of set $\mathcal{A}$ is denoted by $|\mathcal{A}|$. Define the distance metric $D(\boldsymbol{z},\boldsymbol{z}') \triangleq \|\boldsymbol{z}-\boldsymbol{z}'\|$, and the point to set distance as $D(\boldsymbol{z},\mathcal{Z}) \triangleq \inf_{\boldsymbol{z}'\in\mathcal{Z}} D(\boldsymbol{z},\boldsymbol{z}')$. Define $\text{proj}(\boldsymbol{z},\mathcal{Z}) \triangleq \{\boldsymbol{z}' \in \boldsymbol{\mathcal{Z}} | D(\boldsymbol{z},\boldsymbol{z}') = D(\boldsymbol{z},\mathcal{Z})\}$ the projection set of point $\boldsymbol{z}$ onto set $\mathcal{Z}$. Denote the supremum of a function $\eta : \boldsymbol{\mathcal{Z}} \to \mathbb{R}$ as $\|\eta\|_\infty \triangleq \sup_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} |\eta(\boldsymbol{z})|$. For a real number a, we denote the floor function by $\lfloor a \rfloor$. We denote by $\mathcal{O}(g(t))$ the limiting behavior of some function $f(t)$ if $\lim_{t\to\infty} \frac{f(t)}{g(t)} = a$ for some constant $a > 0$.

We give formal definitions of sub-Gaussian random variables and Gaussian process.

Definition 1 *[20] A random variable X with mean $\mu \triangleq \mathbb{E}[X]$ is sub-Gaussian (denoted by $X \sim subG(\sigma^2)$) if there exists a positive number σ such that*

$$\mathbb{E}\left[\exp(\lambda(X-\mu))\right] \le \exp\left(\frac{\sigma^2\lambda^2}{2}\right), \quad \forall \lambda \in \mathbb{R}. \tag{1}$$

Definition 2 *[13] A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

1 Preliminary

In this section, we provide basic knowledge of GPR. Let $\eta : \boldsymbol{\mathcal{Z}} \to \boldsymbol{\mathcal{Y}}$ be the target function, where $\boldsymbol{\mathcal{Z}} \subseteq \mathbb{R}^{n_z}$ and $\boldsymbol{\mathcal{Y}} \subseteq \mathbb{R}$. Given input $\boldsymbol{z}(t) \in \boldsymbol{\mathcal{Z}}$ at time t, the corresponding output is given by

$$y(t) = \eta(\boldsymbol{z}(t)) + e(t), \quad e(t) \sim \mathcal{N}(0, \sigma_e^2), \tag{2}$$

where $e(t)$ is the Gaussian measurement noise and the latent function η is sampled from a Gaussian process prior. Let training data be in the form $\mathcal{D} \triangleq (\mathcal{Z}, \boldsymbol{y})$, where $\mathcal{Z} \triangleq \{\boldsymbol{z}(1), \ldots, \boldsymbol{z}(n_s)\}$ is the set of input data and $\boldsymbol{y} \triangleq [y(1), \ldots, y(n_s)]^{\mathrm{T}}$ is the column vector aggregating the outputs. GPR aims to approximate the latent function η over a set of test data points $\mathcal{Z}_* \subset \boldsymbol{\mathcal{Z}}$ using $\mathcal{D}$.

Define kernel function $k : \mathbb{R}^{n_z} \times \mathbb{R}^{n_z} \to \mathbb{R}$ that is symmetric and positive semi-definite; i.e., $\int k(\boldsymbol{z}, \boldsymbol{z}')f(\boldsymbol{z})f(\boldsymbol{z}')d\nu(\boldsymbol{z})d\nu(\boldsymbol{z}') \geq 0$ for all $f \in \mathcal{L}_2(\boldsymbol{\mathcal{Z}}, \nu)$, where ν denotes a measure (see page 80, [13]). Let $\eta(\mathcal{Z})$ return a column vector such that the i-th entry of $\eta(\mathcal{Z})$ equals $\eta(\boldsymbol{z}(i))$. Assume that the latent function η is sampled from a GP prior specified by mean function μ and kernel function k. Then the training outputs $\boldsymbol{y}$ and the test outputs $\boldsymbol{\eta}(\mathcal{Z}_*)$ are jointly Gaussian distributed as:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{\eta}(\mathcal{Z}_*) \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} \boldsymbol{\mu}(\mathcal{Z}) \\ \boldsymbol{\mu}(\mathcal{Z}_*) \end{bmatrix}, \begin{bmatrix} k(\mathcal{Z}, \mathcal{Z}) + \sigma_e^2 I_{n_s} & k(\mathcal{Z}, \mathcal{Z}_*) \\ k(\mathcal{Z}_*, \mathcal{Z}) & k(\mathcal{Z}_*, \mathcal{Z}_*) \end{bmatrix}),$$

where $\boldsymbol{\mu}(\mathcal{Z})$ and $\boldsymbol{\mu}(\mathcal{Z}_*)$ return a vector consisting of $\mu(\boldsymbol{z}(i))$ and $\mu(\boldsymbol{z}_*(j))$; $k(\mathcal{Z}, \mathcal{Z}_*)$ returns a matrix such that the entry at the ith row and the jth column is $k(\boldsymbol{z}(i), \boldsymbol{z}_*(j))$, and $k(\mathcal{Z}, \mathcal{Z})$, $k(\mathcal{Z}_*, \mathcal{Z})$ and $k(\mathcal{Z}_*, \mathcal{Z}_*)$ are defined in an analogous way.

Utilizing identities of joint Gaussian distribution (see page 200 in [13]), GPR makes predictions of $\boldsymbol{\eta}$ on $\mathcal{Z}_*$ based on dataset $\mathcal{D}$ as $\boldsymbol{\eta}(\mathcal{Z}_*) \sim \mathcal{N}(\boldsymbol{\mu}_{\mathcal{Z}_*|\mathcal{D}}, \Sigma_{\mathcal{Z}_*|\mathcal{D}})$, where

$$\begin{aligned} \boldsymbol{\mu}_{\mathcal{Z}_*|\mathcal{D}} &\triangleq \mu(\mathcal{Z}_*) + k(\mathcal{Z}_*, \mathcal{Z})\mathring{k}(\mathcal{Z}, \mathcal{Z})^{-1}(\boldsymbol{y} - \mu(\mathcal{Z})), \\ \Sigma_{\mathcal{Z}_*|\mathcal{D}} &\triangleq k(\mathcal{Z}_*, \mathcal{Z}_*) - k(\mathcal{Z}_*, \mathcal{Z})\mathring{k}(\mathcal{Z}, \mathcal{Z})^{-1}k(\mathcal{Z}, \mathcal{Z}_*), \end{aligned} \tag{3}$$

with $\mathring{k}(\mathcal{Z}, \mathcal{Z}) \triangleq k(\mathcal{Z}, \mathcal{Z}) + \sigma_e^2 I_{n_s}$. We refer (3) as full GPR. If y is multi-dimensional, GPR is performed for each element.

2 Problem formulation

2.1 Observation model

Consider a network consisting of n agents and a cloud. We denote the agent set by $\mathcal{V} \triangleq \{1, \ldots, n\}$. At each time instant t, the observation model of agent i is given as

$$y^{[i]}(t) = \eta(\boldsymbol{z}^{[i]}(t)) + e^{[i]}(t) \tag{4}$$

where $\boldsymbol{z}^{[i]}(t) \in \boldsymbol{\mathcal{Z}} \subseteq \mathbb{R}^{n_z}$ is the input of η, $y^{[i]}(t) \in \boldsymbol{\mathcal{Y}}$ is the observation of the agent, and $e^{[i]}(t)$ is independent zero-mean Gaussian noise with $e^{[i]}(t) \sim \mathcal{N}(0, (\sigma_e^{[i]})^2)$.

2.2 Attack model

All communications between the cloud and the agents are attack-free. Some agents are Byzantine agents, which can behave arbitrarily during learning process, and send arbitrary local predictions to the cloud. We assume that an α fraction of the n agents are Byzantine, and the remaining $1 - \alpha$ fraction are benign. We denote $\mathcal{B} \subset \mathcal{V}$ as the set of Byzantine agents, and thus $|\mathcal{B}| = \lfloor \alpha n \rfloor$. Both the cloud and the benign agents are unaware of the identities of Byzantine agents, but we assume the cloud knows that $\frac{1}{4}$ is an upper bound of α.

Assumption 1 *It holds that* $0 < \alpha < \frac{1}{4}$.

Most existing works on Byzantine-tolerant algorithms, e.g., [5], [7], [9], can tolerate Byzantine agents with $\alpha < \frac{1}{2}$. These algorithms only transmit one variable, e.g., the gradients of the loss function or the model parameter, between the cloud and the agents, whereas our federated GPR algorithm requires to exchange two variables, e.g., local mean and variance, to realize the aggregation of predictions.

2.3 Objective

The objective of this paper is to design an algorithm, which enables the benign agents and the cloud to correctly learn the function η without requiring the agents to share local streaming data $(\boldsymbol{z}^{[i]}(t), y^{[i]}(t))$ despite some agents are subject to Byzantine attacks.

3 Byzantine-tolerant federated GPR

In this section, we develop a Byzantine-tolerant federated GPR algorithm to solve the problem formulated in Section 2. This algorithm consists of three modules: agent-based local GPR, cloud-based aggregated GPR and agent-based fused GPR.

Algorithm 1 Byzantine-tolerant federated GPR

Input: test inputs $\mathcal{Z}_*$, prior mean function μ, kernel k and noise variance $(\sigma_e^{[i]})^2$ for $i \in \mathcal{V}$
Initialization: $\mathcal{D}(0) = \emptyset$.
for $t = 1, 2, \ldots$ **do**
  for $i \in \mathcal{V}$ **do** {Agent-based local GPR}
    $\mathcal{D}^{[i]}(t) = \mathcal{D}^{[i]}(t-1) \bigcup (\boldsymbol{z}^{[i]}(t), y^{[i]}(t))$
    $\check{\boldsymbol{\mu}}'_{\mathcal{Z}_*|\mathcal{D}^{[i]}(t)}, \check{\boldsymbol{\sigma}}'^2_{\mathcal{Z}_*|\mathcal{D}^{[i]}(t)} = \text{lGPR}(\mathcal{D}^{[i]}(t))$
  end for
  $\hat{\boldsymbol{\mu}}_{\mathcal{Z}_*|\mathcal{D}(t)}, \hat{\boldsymbol{\sigma}}^2_{\mathcal{Z}_*|\mathcal{D}(t)} = \text{cGPR}(\check{\boldsymbol{\mu}}'_{\mathcal{Z}_*|\mathcal{D}^{[i]}(t)}, \check{\boldsymbol{\sigma}}'^2_{\mathcal{Z}_*|\mathcal{D}^{[i]}(t)})$ {Cloud-based aggregated GPR}
  for $i \in \mathcal{V}$ **do** {Agent-based fused GPR}
    $\tilde{\boldsymbol{\mu}}^{[i]}_{\mathcal{Z}_*|\mathcal{D}(t)}, (\tilde{\boldsymbol{\sigma}}^{[i]}_{\mathcal{Z}_*|\mathcal{D}(t)})^2 = \text{fGPR}(\check{\boldsymbol{\mu}}'_{\mathcal{Z}_*|\mathcal{D}^{[i]}(t)}, \check{\boldsymbol{\sigma}}'^2_{\mathcal{Z}_*|\mathcal{D}^{[i]}(t)}, \hat{\boldsymbol{\mu}}_{\mathcal{Z}_*|\mathcal{D}(t)}, \hat{\boldsymbol{\sigma}}^2_{\mathcal{Z}_*|\mathcal{D}(t)})$
  end for
end for

Specifically, first at each time instant t, each agent $i \in \mathcal{V}$ makes its own prediction of η through agent-based local GPR over a given set of points $\mathcal{Z}_* \subseteq \boldsymbol{\mathcal{Z}}$ using local streaming dataset $\mathcal{D}^{[i]}(t) \triangleq (\mathcal{Z}^{[i]}(t), \boldsymbol{y}^{[i]}(t))$ with local input data $\mathcal{Z}^{[i]}(t) \triangleq \{\boldsymbol{z}^{[i]}(1), \ldots, \boldsymbol{z}^{[i]}(t)\}$ and output $\boldsymbol{y}^{[i]}(t) \triangleq [y^{[i]}(1), \ldots, y^{[i]}(t)]^{\mathrm{T}}$. Then the benign agents send correct local predictions to the cloud and the Byzantine agents instead send arbitrary values. Second, the cloud collects all the local predictions and constructs two agent sets by trimming a fraction of outliers with respect to the predictive means and variances. Then the cloud uses a Byzantine-tolerant PoE aggregation rule to compute a global prediction, and broadcasts the global prediction to all the agents. Third, the agents refine the predictions on $\mathcal{Z}_*$ by fusing the prediction from the cloud-based GPR with that from the agent-based local GPR. Note that each agent only communicates local predictions with the cloud, and does not share its local streaming data with the cloud and other agents. The formal description of Byzantine-tolerant federated GPR is given in Algorithm 1.

Algorithm 2 Agent-based local GPR: $\text{lGPR}(\mathcal{D}^{[i]}(t))$

$\mathcal{Z}^{[i]}(t) = \{\boldsymbol{z}^{[i]}(1), \ldots, \boldsymbol{z}^{[i]}(t)\}$
$\boldsymbol{y}^{[i]}(t) = [y^{[i]}(1), \ldots, y^{[i]}(t)]^{\mathrm{T}}$
for $\boldsymbol{z}_* \in \mathcal{Z}_*$ **do**
  Choose $z_*^{[i]}(t) \in \text{proj}(\boldsymbol{z}_*, \mathcal{Z}^{[i]}(t))$
  Compute $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \leftarrow \begin{cases} \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} & \text{Benign agent,} \\ \star & \text{Byzantine agent} \end{cases}$
  Compute $\check{\sigma}'^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \leftarrow \begin{cases} \check{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} & \text{Benign agent,} \\ \star & \text{Byzantine agent} \end{cases}$
  Send $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ and $\check{\sigma}'^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ to the cloud
end for
return $\check{\boldsymbol{\mu}}'_{\mathcal{Z}_*|\mathcal{D}^{[i]}(t)}, \check{\boldsymbol{\sigma}}'^2_{\mathcal{Z}_*|\mathcal{D}^{[i]}(t)}$

3.1 Agent-based local GPR

To reduce computational complexity, we perform Nearest-neighbor (NN) GPR as the agent-based local GPR to predict for each test point $\boldsymbol{z}_* \in \mathcal{Z}_*$. Instead of feeding the whole training dataset to GPR, the agent-based local GPR only uses the nearest input denoted by $\boldsymbol{z}_*^{[i]}(t) \in \text{proj}(\boldsymbol{z}_*, \mathcal{Z}^{[i]}(t))$ and its corresponding output $y^{[i]}_{\boldsymbol{z}_*^{[i]}(t)}$ to compute the local predictions. Notice that, computation complexity of the hyperparameter tuning for NN GPR is $\mathcal{O}(t^3)$. In fact, the computation complexity of the hyperparameter tuning can be reduced to $\mathcal{O}(1)$ by using recursive hyperparameter tuning [21]. NN GPR in prediction has computation complexity equivalent to nearest-neighbor search, whose worst case complexity is $\mathcal{O}(t)$ [22], while full GPR in (3) has computation complexity $\mathcal{O}(t^3)$ [23]. Denote $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ and $\check{\sigma}'^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ the predictive mean and variance output from the agent-based local GPR. The predictive mean and variance of the benign agents are written in the following forms

$\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} = \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$, $\check{\sigma}'^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} = \check{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ where

$$\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \triangleq \mu(\boldsymbol{z}_*) + k(\boldsymbol{z}_*, \boldsymbol{z}_*^{[i]}(t))\mathring{k}(\boldsymbol{z}_*^{[i]}(t), \boldsymbol{z}_*^{[i]}(t))^{-1}(y^{[i]}_{\boldsymbol{z}_*^{[i]}(t)} - \mu(\boldsymbol{z}_*^{[i]}(t))),$$
$$\check{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \triangleq k(\boldsymbol{z}_*, \boldsymbol{z}_*) - k(\boldsymbol{z}_*, \boldsymbol{z}_*^{[i]}(t))\mathring{k}(\boldsymbol{z}_*^{[i]}(t), \boldsymbol{z}_*^{[i]}(t))^{-1}k(\boldsymbol{z}_*^{[i]}(t), \boldsymbol{z}_*) \tag{5}$$

with $\mathring{k}(\boldsymbol{z}_*^{[i]}(t), \boldsymbol{z}_*^{[i]}(t)) \triangleq k(\boldsymbol{z}_*^{[i]}(t), \boldsymbol{z}_*^{[i]}(t)) + (\sigma_e^{[i]})^2 I_t$. When the agent is a Byzantine agent, its outputs can be arbitrary, and in this case we write $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} = \star$ and $\check{\sigma}'^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} = \star$. The implementation of the agent-based local GPR is provided in Algorithm 2.

3.2 Cloud-based aggregated GPR

In the sequel, we consider a Byzantine-tolerant PoE aggregation rule for the cloud. The algorithm is described as follows:

Step 1: For each $\boldsymbol{z}_* \in \mathcal{Z}_*$, the cloud collects potentially corrupted mean $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ and variance $\check{\sigma}'^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ from each agent i.

Step 2: The cloud constructs two agent sets to tolerate Byzantine attacks. Without loss of generality, for $i \in \mathcal{V}$, we sort $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ and $\check{\sigma}'^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ in ascending order and the corresponding agent index is kept in the sets $\mathcal{T}^{\mu}(t)$ and $\mathcal{T}^{\sigma}(t)$, such that for each $i \in \mathcal{T}^{\mu}(t)$, we have $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \leq \check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i+1]}(t)}$ and for each $i \in \mathcal{T}^{\sigma}(t)$, we have $\check{\sigma}'^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \leq \check{\sigma}'^2_{\boldsymbol{z}_*|\mathcal{D}^{[i+1]}(t)}$. We denote by $\mathcal{T}^{\mu}_{\max}(t) \triangleq \{n - \lfloor \beta n \rfloor + 1, \ldots, n\}$ a set of $\lfloor \beta n \rfloor$ agents with the largest local predictive means, and by $\mathcal{T}^{\mu}_{\min}(t) \triangleq \{1, \ldots, \lfloor \beta n \rfloor\}$ a set of $\lfloor \beta n \rfloor$ agents with the smallest local predictive means where $\beta \in [\alpha, \frac{1}{4})$. Similarly, we denote trimmed sets of variance by $\mathcal{T}^{\sigma}_{\max}(t)$ and $\mathcal{T}^{\sigma}_{\min}(t)$. Then, we define

$$\mathcal{M}(t) \triangleq \mathcal{V} \backslash (\mathcal{T}^{\mu}_{\max}(t) \bigcup \mathcal{T}^{\mu}_{\min}(t)), \quad \mathcal{C}(t) \triangleq \mathcal{V} \backslash (\mathcal{T}^{\sigma}_{\max}(t) \bigcup \mathcal{T}^{\sigma}_{\min}(t)). \tag{6}$$

Notice that $\mathcal{M}(t)$ and $\mathcal{C}(t)$ include $n - 2\lfloor \beta n \rfloor$ agents with extreme values removed. These are the agents whose predictions do not deviate much from the attack-free case.

Step 3: Define a global training dataset as $\mathcal{D}(t) \triangleq \bigcup_{i \in \mathcal{V}} \mathcal{D}^{[i]}(t)$, and a common set as $\mathcal{I}(t) \triangleq \mathcal{M}(t) \bigcap \mathcal{C}(t)$. The following lemma shows that $\mathcal{I}(t)$ is never empty.

Lemma 1 *Suppose that $\beta < \frac{1}{4}$, it holds that $(1 - 4\beta)n \leq |\mathcal{I}(t)| \leq (1 - 2\beta)n$ at each time instant t.*

The proof of Lemma 1 can be found in Appendix A.1. The following lemma shows that $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ and $\check{\sigma}'^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ for $i \in \mathcal{I}(t)$ are bounded by the values of the benign agents.

Algorithm 3 Cloud-based aggregated GPR: cGPR($\check{\boldsymbol{\mu}}'_{\mathcal{Z}_*|\mathcal{D}^{[i]}(t)}, \check{\boldsymbol{\sigma}}'^2_{\mathcal{Z}_*|\mathcal{D}^{[i]}(t)}$)

```
for t = 1, 2 ... do
    D(t) = ⋃_{i∈V} D^[i](t)
    for z_* ∈ Z_* do
        Construct sets M(t) and C(t) according to (6)
        Compute I(t) = M(t) ⋂ C(t)
        Implement aggregation rule (7)
        Broadcast μ̂_{Z_*|D(t)} and σ̂²_{Z_*|D(t)} to all the agents
    end for
end for
return μ̂_{Z_*|D(t)}, σ̂²_{Z_*|D(t)}
```

Lemma 2 *Let $0 < \alpha \leq \beta < \frac{1}{4}$ hold. For $i \in \mathcal{I}(t)$, we have that $\min_{j \in \mathcal{V} \backslash \mathcal{B}} \left\{\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)}\right\} \leq \check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \leq \max_{j \in \mathcal{V} \backslash \mathcal{B}} \left\{\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)}\right\}$, and $\min_{j \in \mathcal{V} \backslash \mathcal{B}} \left\{\check{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)}\right\} \leq \check{\sigma}'^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \leq \max_{j \in \mathcal{V} \backslash \mathcal{B}} \left\{\check{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)}\right\}$.*

The proof of Lemma 2 is presented in Appendix A.2. Given $\mathcal{I}(t)$, we propose a Byzantine-tolerant PoE aggregation rule below

$$\hat{\mu}_{\boldsymbol{z}_*|\mathcal{D}(t)} \triangleq \frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|} \sum_{i\in\mathcal{I}(t)} \check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}, \quad \hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)} \triangleq \frac{|\mathcal{I}(t)|}{\sum_{i\in\mathcal{I}(t)} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}}. \tag{7}$$

Note that if all the agents are known to be benign, we can simply select $\beta = \alpha = 0$, and hence $\mathcal{M}(t) = \mathcal{C}(t) = \mathcal{V}$. In this case, (7) reduces to the standard PoE in [24]. Then the cloud-based aggregated GPR returns $\hat{\boldsymbol{\mu}}_{\mathcal{Z}_*|\mathcal{D}(t)} \triangleq [\hat{\mu}_{\boldsymbol{z}_*|\mathcal{D}(t)}]_{\boldsymbol{z}_*\in\mathcal{Z}_*}$ and $\hat{\boldsymbol{\sigma}}^2_{\mathcal{Z}_*|\mathcal{D}(t)} \triangleq [\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}]_{\boldsymbol{z}_*\in\mathcal{Z}_*}$. The implementation of the cloud-based aggregated GPR is provided in Algorithm 3.

3.3 Agent-based fused GPR

Notice that the cloud makes predictions using more data than each local agent, and it could achieve better learning performances. Then the agent-based fused GPR aims to refine predictions of $\eta(\mathcal{Z}_*)$ by integrating the agent-based local GPR with the cloud-based aggregated GPR. When the cloud variance $\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}$ is lower than the local predictive variances $\check{\sigma}'^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$, the predictions from the cloud-based aggregated GPR are used. Otherwise, those from the agent-based local GPR are used. Our algorithm is one-round. The refined predictions in the agent-based fused GPR will not be transmitted back to the cloud, hence it cannot change the final aggregation. The implementation of the agent-based fused GPR is provided in Algorithm 4.

Algorithm 4 Agent-based fused GPR: fGPR$(\check{\boldsymbol{\mu}}'_{\mathcal{Z}_*|\mathcal{D}^{[i]}(t)}, \check{\boldsymbol{\sigma}}'^2_{\mathcal{Z}_*|\mathcal{D}^{[i]}(t)}, \hat{\boldsymbol{\mu}}_{\mathcal{Z}_*|\mathcal{D}(t)}, \hat{\boldsymbol{\sigma}}^2_{\mathcal{Z}_*|\mathcal{D}(t)})$

```
for z_* ∈ Z_* do
    if σ̂²_{z_*|D(t)} ≤ σ̌'²_{z_*|D^[i](t)} then
        μ̃^[i]_{z_*|D(t)} = μ̂_{z_*|D(t)},   (σ̃^[i]_{z_*|D(t)})² = σ̂²_{z_*|D(t)}
    else
        μ̃^[i]_{z_*|D(t)} = μ̌'_{z_*|D^[i](t)},   (σ̃^[i]_{z_*|D(t)})² = σ̌'²_{z_*|D^[i](t)}
    end if
end for
return μ̃^[i]_{Z_*|D(t)}, (σ̃^[i]_{Z_*|D(t)})²
```

3.4 Performance guarantee

This section discusses the robustness of the cloud-based aggregated GPR. We start with the assumptions of the kernel function k.

Assumption 2 *1) (Decomposition). The kernel function $k(\cdot,\cdot)$ can be decomposed in such a way that $k(\cdot,\cdot) = \kappa(D(\cdot,\cdot))$, where $\kappa : \mathbb{R}_+ \to \mathbb{R}_{++}$ is continuous.*

2) (Monotonicity). It holds that $\kappa(s)$ is monotonically decreasing in s and $\kappa(0) = \sigma_f^2$.

Many existing kernel functions can satisfy Assumption 2, e.g., squared exponential kernel [13]. As [13], any non-zero mean Gaussian process can be decomposed into a deterministic process and a zero-mean stochastic process. Without loss of generality and for notational simplicity, we assume η follows a zero-mean Gaussian process in the analysis.

Assumption 3 *It satisfies that $\eta \sim \mathcal{GP}(0, k)$ and η is Lipschitz continuous on $\boldsymbol{\mathcal{Z}}$.*

The target function η is completely specified by Gaussian process with kernel function k. This assumption is common in the analysis of GPR [25]. The Lipschitz continuity of η implies that there exists some positive constant ℓ_η such that $|\eta(\boldsymbol{z}) - \eta(\boldsymbol{z}')| \leq \ell_\eta D(\boldsymbol{z}, \boldsymbol{z}')$. For $i \in \mathcal{V}$, we define the dispersion of local data as $d^{[i]}(t) \triangleq \sup_{\boldsymbol{z}\in\boldsymbol{\mathcal{Z}}} D(\boldsymbol{z}, \mathcal{Z}^{[i]}(t))$. For notational simplicity, we denote the distance by $D^{\mathcal{Z}}_{\boldsymbol{z}} \triangleq D(\boldsymbol{z}, \mathcal{Z})$.

We denote $\sigma \triangleq \frac{\sigma_f^2 \sigma_e^{\max}}{\sigma_f^2+(\sigma_e^{\min})^2}$ where $\sigma_e^{\max} \triangleq \max_{i\in\mathcal{V}}\{\sigma_e^{[i]}\}$ and $\sigma_e^{\min} \triangleq \min_{i\in\mathcal{V}}\{\sigma_e^{[i]}\}$. The following theorem shows that the prediction error from the cloud-based aggregated GPR is upper bounded, and characterizes the bounds of the global predictive variance.

Theorem 1 ***Part I:*** *(Cloud-based aggregated GPR: Mean) Suppose $0 < \alpha \leq \beta < \frac{1}{4}$ and Assumptions 2, 3 hold. For any $\boldsymbol{z}_* \in \boldsymbol{\mathcal{Z}}$ and $0 < \delta < 1$, with probability at least $1-\delta$, it holds that $\left|\hat{\mu}_{\boldsymbol{z}_*|\mathcal{D}(t)} - \eta(\boldsymbol{z}_*)\right| \leq (1 - \frac{\kappa(d^{\max}(t))}{\sigma_f^2+(\sigma_e^{\max})^2})\|\eta\|_\infty + \frac{\sigma_f^2 \ell_\eta d^{\max}(t)}{\sigma_f^2+(\sigma_e^{\min})^2} + \sqrt{2\sigma^2(\ln 2 - \ln\delta)} + \Delta(d^{\max}(t))$ where $\Delta(s) \triangleq \frac{2\alpha(\sqrt{2\sigma^2(\ln(2n)-\ln\delta)}+\frac{\sigma_f^2\|\eta\|_\infty}{\sigma_f^2+(\sigma_e^{\min})^2})}{1-4\beta}\frac{\sigma_f^4+\sigma_f^2(\sigma_e^{\max})^2-\kappa(s)^2}{\sigma_f^2(\sigma_e^{\min})^2}$.*

Part II: *(Cloud-based aggregated GPR: Variance) Let $0 < \alpha \leq \beta < \frac{1}{4}$ and Assumption 2 hold. For any $\boldsymbol{z}_* \in \boldsymbol{\mathcal{Z}}$, it holds that $\frac{\sigma_f^2(\sigma_e^{\min})^2}{\sigma_f^2+(\sigma_e^{\max})^2} \leq \hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)} \leq \sigma_f^2 - \frac{\kappa(d^{\max}(t))^2}{\sigma_f^2+(\sigma_e^{\max})^2}$.*

The proof of Theorem 1 is presented in Appendix A.3. For the attack-free case, i.e., $\alpha = \beta = 0$, the upper bound of Part I becomes $(1-\frac{\kappa(d^{\max}(t))}{\sigma_f^2+(\sigma_e^{\max})^2})\|\eta\|_\infty + \frac{\sigma_f^2\ell_\eta d^{\max}(t)}{\sigma_f^2+(\sigma_e^{\min})^2} + \sqrt{2\sigma^2(\ln 2 - \ln\delta)}$. We assume $\lim_{t\to\infty} d^{[i]}(t) = 0$ for $i \in \mathcal{V}$, the limit of the upper bound is $\frac{(\sigma_e^{\max})^2}{\sigma_f^2+(\sigma_e^{\max})^2}\|\eta\|_\infty + \sqrt{2\sigma^2(\ln 2 - \ln\delta)}$. Recall that $\sigma = \frac{\sigma_f^2\sigma_e^{\max}}{\sigma_f^2+(\sigma_e^{\min})^2}$. Then we can see that $\frac{(\sigma_e^{\max})^2}{\sigma_f^2+(\sigma_e^{\max})^2}\|\eta\|_\infty$ and $\sqrt{2\sigma^2(\ln 2 - \ln\delta)}$ are positively related to the maximal variance $(\sigma_e^{\max})^2$ of the noise.

The term $\Delta(d^{\max}(t))$ characterizes the performance loss caused by Byzatine agents. First, notice that $\Delta(d^{\max}(t))$ is positively related to α, which provides a trade-off between the prediction error and the number of Byzantine agents. With a larger α, our algorithm can tolerate more Byzantine agents, but the prediction error gets larger. As α approaches zero, $\Delta(d^{\max}(t))$ goes to zero. This implies that the prediction performance improves if the number of Byzantine agents is decreasing. Second, when α is fixed, $\Delta(d^{\max}(t))$ is proportional to β. This implies that learning performance improves if the set $\mathcal{I}(t)$ expands.

The upper and lower bounds in Part II are independent of α and β. Then the variance shares the same upper and lower bounds with that in the attack-free case. This is due to the fact that the extreme values are removed by the Byzantine-tolerant PoE in the cloud-based GPR. As t goes to infinity, we have $\limsup_{t\to\infty} \hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)} \leq \frac{\sigma_f^2(\sigma_e^{\max})^2}{\sigma_f^2+(\sigma_e^{\max})^2}$. This implies that the prediction uncertainty diminishes as the variance of measurement noise decreases.

4 Numerical experiments

This section experimentally demonstrates the performance of the proposed Byzantine-tolerant federated GPR algorithm using one synthetic dataset and two real-world datasets. We conduct the experiments on a computer with Intel i7-6600 CPU, 2.60GHz and 12 GB RAM.

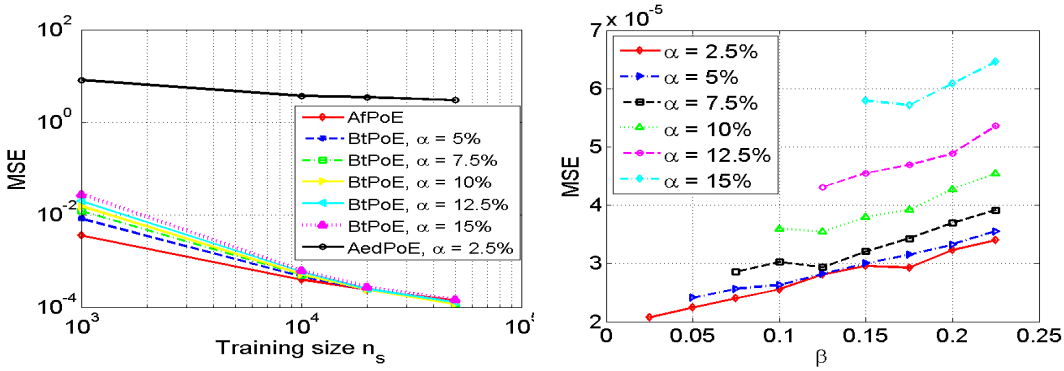
Synthetic dataset: For the observation model (4), we consider the target function introduced in [23]

$$\eta(\boldsymbol{z}) = 5\boldsymbol{z}^2\sin(12\boldsymbol{z}) + (\boldsymbol{z}^3 - 0.5)\sin(3\boldsymbol{z} - 0.5) + 4\cos(2\boldsymbol{z}), \tag{8}$$

with $\boldsymbol{z} \in \mathbb{R}$ and $e \sim \mathcal{N}(0, 0.01)$. We generate $n_s = 10^3, 5\times10^3, 10^4, 5\times10^4$ training points in $[0,1]$, respectively, and choose $n_t = 120$ test points randomly in $[0,1]$. There are $n = 40$ agents in a network. We partition the training dataset into 40 disjoint groups, and each agent is assigned with n_s/n training data points. We use the following squared-exponential kernel $k(\boldsymbol{z}, \boldsymbol{z}_*) = \sigma_f^2\exp(-\frac{1}{2\ell^2}(\boldsymbol{z} - \boldsymbol{z}_*)^2)$, and use Mean Square Error (MSE) to evaluate the proposed algorithm in terms of consistency, the effect of α and the effect of β in the following experiments.

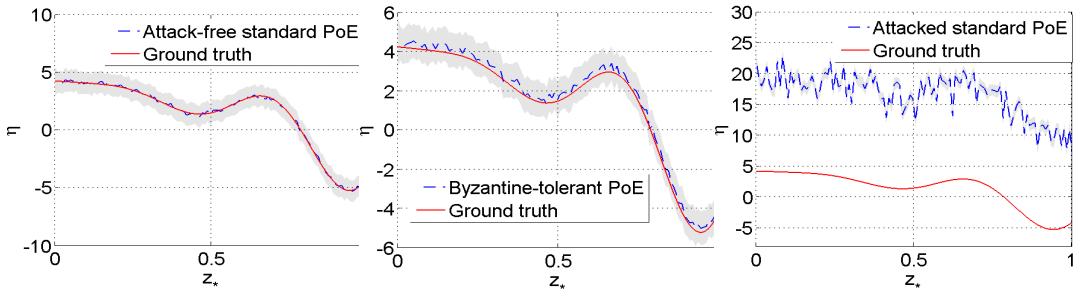
In the first experiment, we let $\alpha = 2.5\%, 5\%, 7.5\%, 10\%, 12.5\%, 15\%$. We randomly choose the agents in the network to be compromised by same-value attacks [26], and let $\beta = \alpha$. Specifically, for each test point $\boldsymbol{z}_*$, the Byzantine agents only change the local predictive means to 100, that is, $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} = 100$ for all t, and send this incorrect prediction to the cloud. The local predictive variances for all the agents remain unchanged. Fig. 1(a) compares the performances of the proposed Byzantine-tolerant PoE (BtPoE) with the standard PoE. As a benchmark, the bottom curve

depicts the prediction errors of the standard PoE (AfPoE) when there is no Byzantine attack in the network; it shows that the standard PoE is consistent in the attack-free scenario. However, when the aforementioned Byzantine attacks are launched, the top curve shows that the standard PoE can no longer remain consistent. Note that there are 6 curves by attacked standard PoE (AedPoE), and only the closest one to the bottom curve is shown in Fig. 1(a). We fix $n_s = 4 \times 10^4$ and $\alpha = 15\%$. The prediction comparisons of these three PoEs are given in Fig. 2, which shows that the learning performance of the Byzantine-tolerant PoE is comparable to that of the attack-free standard PoE with consistency preserved. As Byzantine attacks are launched in the network, the consistency cannot be maintained. Therefore, our proposed algorithm is demonstrated to be resilient to Byzantine attacks.



(a) Consistency evaluation (b) Prediction performance on different β

Figure 1: Prediction performance of the cloud-based aggregated GPR.

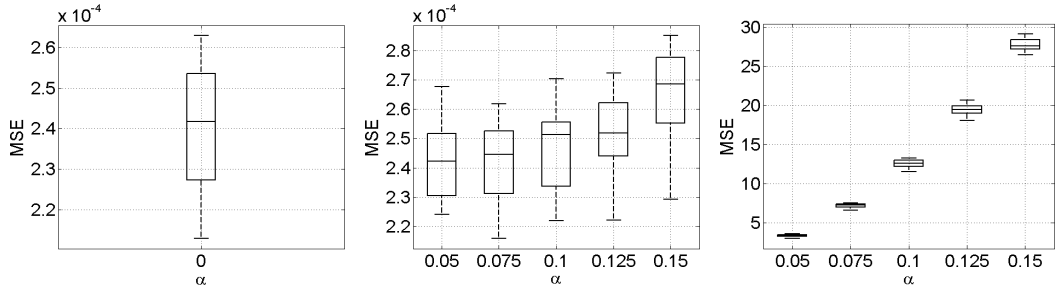


(a) Attack-free standard PoE (b) Byzantine-tolerant PoE (c) Attacked standard PoE

Figure 2: Comparisons of the prediction performance in terms of different PoEs. Shaded area represents 99% confidence region.

Notice that the performance of the Byzantine-tolerant PoE increases as α, the ratio of Byzantine agents, decreases. Fixing the training data size $n_s = 10^4$, we replicate the experiments for 50 times with different α and observation noise e. Figs. 3(a) and 3(b) show that the performances of the Byzantine-tolerant PoE are comparable to that of the attack-free standard PoE with consistency preserved. The prediction errors for both methods are in the order of 10^{-4}, while Fig. 3(c) shows that the attacked standard PoE has prediction errors ranged from 2.5 to 30. This demonstrates that the proposed PoE aggregation rule has strong ability to tolerate Byzantine attacks.

In the second experiment, we evaluate the prediction performance with regard to β. All associated parameters are the same as those in the above experiments. Notice that for each α, we initially let $\beta = \alpha$ and then increase β to 0.225 with increment 0.025, such that the assumption $\alpha \leq \beta < \frac{1}{4}$ is satisfied. The results are shown in Fig. 1(b). This shows that as the number of trimmed agents increases, the performance of the proposed algorithm grows worse, since more information is excluded in the aggregation. This result validates the conclusion of part I in Theorem 1.



(a) Attack-free standard PoE (b) Byzantine-tolerant PoE (c) Attacked standard PoE

Figure 3: Comparison of the prediction performance on different α.

Table 1: Comparison of the prediction performance on different functions.

Algorithm	Attack-free standard PoE	Byzantine-tolerant PoE	Attacked standard PoE
MSE	$0.0049 \pm 0.007 \times 10^{-3}$	$0.0236 \pm 0.172 \times 10^{-3}$	$26.5339 \pm 0.019 \times 10^{-3}$

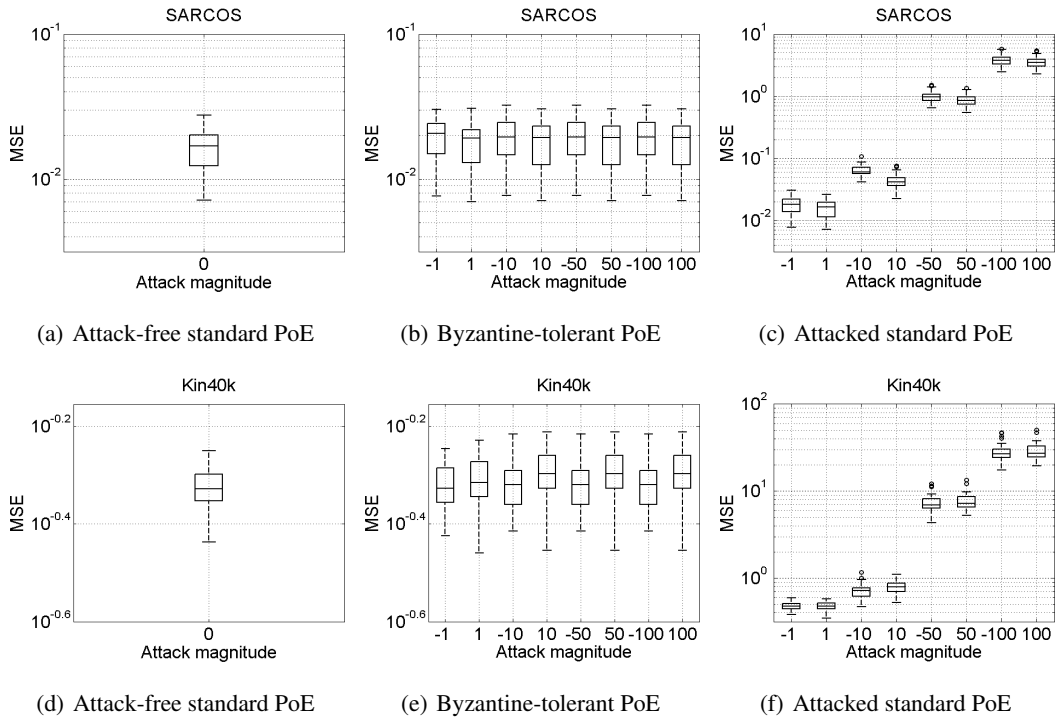
In the third experiment, we conduct Monte Carlo simulations to evaluate the performance of Byzantine-tolerant PoE on different functions. We repeat the experiment for 50 times by sampling different functions and observation points. Specifically, we replace $\sin(12\boldsymbol{z})$ in (8) with $\sin(12\boldsymbol{z}) + \epsilon'$ where $\epsilon' \sim \mathcal{N}(0, 0.01i)$, $i = 1, \ldots, 50$. We assign $n = 40$ agents in the network and sample $n_s = 10^4$ training points in $[0, 1]$, and then choose $n_t = 120$ test points randomly in $[0, 1]$. We also use the square-exponential kernel $k(\boldsymbol{z}, \boldsymbol{z}_*)$. There are six Byzantine agents, i.e., $\alpha = 15\%$, and we let $\beta = \alpha$. Experiment results are presented in Table 1, and it shows that the performance of the Byzantine-tolerant PoE is comparable to that of the attack-free standard PoE, noticing that the order of the prediction errors for the Byzantine-tolerant PoE is 10^{-2}. For comparison, the MSE by the attacked standard PoE is about 26.534. This demonstrates that the proposed Byzantine-tolerant aggregation rule is effective for learning different functions.

Table 2: Prediction performance comparisons between local GPR and fused GPR on agent 1 and 6.

Training data size	6000	10000	40000	150000
MSE/$i = 1$/local GPR	0.0721	0.0681	0.0585	0.0527
MSE/$i = 1$/fused GPR	0.0096	0.0681	0.0094	0.0537
MSE/$i = 6$/local GPR	0.0727	0.0566	0.0564	0.0555
MSE/$i = 6$/fused GPR	0.0727	0.0086	0.0564	0.0076

Moreover, we conduct additional experiments to demonstrate the learning performance improvements by applying our Byzantine-tolerant PoE and the agent-based fused GPR to the synthetic dataset. We assign $n = 40$ agents in the network and sample $n_s = 10^4$ training points in $[0, 1]$, and then choose $n_t = 120$ test points randomly in $[0, 1]$. We also use the square-exponential kernel $k(\boldsymbol{z}, \boldsymbol{z}_*)$. There are six Byzantine agents, i.e., $\alpha = 15\%$, and we let $\beta = \alpha$. The prediction performance comparisons between the agent-based local GPR and fused GPR on agent 1 and 6 are listed in Table 2. For a test point, the local variances $\check{\sigma}'^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ lie in $[0.3660, 0.6228]$, and the global variance is $\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)} = 0.4678$. Table 2 shows that the learning performance of the agent-based fused GPR is better than that of the agent-based local GPR. Note that the order of the prediction errors for the agent-based fused GPR can decrease from 10^{-2} to 10^{-3}. This demonstrates that the comparison between variances in step 2 and step 5 in Algorithm 4 is able to improve the learning accuracy, and our algorithm is effective.

Real-world dataset: We conduct an experiment to evaluate the prediction performance with different attack magnitudes using two real-world datasets. The first dataset is collected from a seven degrees-of-freedom SARCOS anthropomorphic robot arm [13]. The goal is to learn a function η. The input of η is 21-dimensional including 7 joint positions, 7 joint velocities and 7 joint accelerations; whereas the output is 1-dimensional, representing the torque of a joint. We randomly sample 40,000 data points for training and 40 data points for testing. In the experiment, we partition $40,000$ training data into $n = 40$ disjoint groups, and assign each group to an agent. The second dataset *Kin40k* [27] is created



(a) Attack-free standard PoE (b) Byzantine-tolerant PoE (c) Attacked standard PoE

(d) Attack-free standard PoE (e) Byzantine-tolerant PoE (f) Attacked standard PoE

Figure 4: Prediction performance with different attack magnitudes on datasets SARCOS and *Kin40k*.

using a robot arm simulator. We randomly choose $9,000$ training points and 40 test points. We also partition the training data evenly into 40 disjoint groups. The magnitudes of attack we consider are $\pm 1, \pm 10, \pm 50, \pm 100$. We randomly sample the training data points and repeat the experiments for 50 times. Fig. 4 shows the prediction performances. For both real-world datasets, when the attack magnitude is increasing, the prediction errors of the attacked standard PoE largely increases, whereas those of the attack-free standard PoE and the Byzantine-tolerant PoE are consistently maintained at a low level. It shows that the proposed Byzantine-tolerant PoE is resilient to the attack.

5 Conclusion

We propose a Byzantine-tolerant federated GPR algorithm, which is able to tolerate less than one quarter Byzantine agents in the network. We derive the upper bounds on the prediction errors and the lower and upper bounds of the predictive variances. Experiments are conducted to demonstrate the robustness of Byzantine-tolerant GPR algorithm. Future work will focus on the comparison of the learning performance between the agent-based local GPR and the agent-based fused GPR, and the quantification of the benefits the cloud brings to the agents.

6 Acknowledgements

This work was partially supported by NSF awards ECCS 1846706 and ECCS 2140175.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

[2] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *Proceedings of International Conference on Neural Information Processing Systems*, pp. 1–5, 2016.

[3] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–19, 2019.

[4] F. Yin, Z. Lin, Q. Kong, Y. Xu, D. Li, S. Theodoridis, and S. R. Cui, "Fedloc: Federated learning framework for data-driven cooperative localization and location data processing," *IEEE Open Journal of Signal Processing*, vol. 1, pp. 187 – 215, 2020.

[5] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proceedings of International Conference on Neural Information Processing Systems*, pp. 118–128, 2017.

[6] J. So, B. Güler, and A. S. Avestimehr, "Byzantine-resilient secure federated learning," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 2168–2181, 2020.

[7] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *Proceedings of the International Conference on Machine Learning*, pp. 5650–5659, 2018.

[8] D. Data and S. Diggavi, "Byzantine-resilient high-dimensional SGD with local iterations on heterogeneous data," in *Proceedings of the International Conference on Machine Learning*, pp. 2478–2488, 2021.

[9] Y. Chen, L. Su, and J. Xu, "Distributed statistical machine learning in adversarial settings: Byzantine gradient descent," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 2, pp. 1–25, 2017.

[10] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[11] L. Bottou, *Online algorithms and stochastic approximations*. Cambridge University Press, 1998.

[12] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, pp. 551–585, 2006.

[13] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

[14] H. Liu, Y.-S. Ong, X. Shen, and J. Cai, "When Gausian process meets big data: A review of scalable GPs," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4405–4423, 2020.

[15] H. Xiao and C. Eckert, "Lazy Gaussian process committee for real-time online regression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 969–976, 2013.

[16] K. Das and A. N. Srivastava, "Sparse inverse kernel Gaussian process regression," *Statistical Analysis and Data Mining*, vol. 6, no. 3, pp. 205–220, 2013.

[17] C. A. Micchelli, Y. Xu, and H. Zhang, "Universal kernels," *Journal of Machine Learning Research*, vol. 7, no. 95, pp. 2651–2667, 2006.

[18] Y. Kuang, T. Chen, F. Yin, and R. Zhong, "Recursive implementation of Gaussian process regression for spatial-temporal data modeling," in *International Conference on Wireless Communications and Signal Processing*, pp. 1–7, 2019.

[19] M. Liu, G. Chowdhary, B. C. Da Silva, S.-Y. Liu, and J. P. How, "Gaussian processes for learning and control: A tutorial with examples," *IEEE Control Systems Magazine*, vol. 38, no. 5, pp. 53–86, 2018.

[20] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. NY Cambridge University Press, 2019.

[21] M. F. Huber, "Recursive Gaussian process: On-line regression and learning," *Pattern Recognition Letters*, vol. 45, pp. 85 – 91, 2004.

[22] M. R. Abbasifard, B. Ghahremani, and H. Naderi, "A survey on nearest neighbor search methods," *International Journal of Computer Applications*, vol. 95, no. 25, 2014.

[23] H. Liu, J. Cai, Y. Wang, and Y. S. Ong, "Generalized robust Bayesian committee machine for large-scale Gaussian process regression," in *Proceedings of the International Conference on Machine Learning*, pp. 3131–3140, 2018.

[24] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[25] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-theoretic regret bounds for Gaussian process optimization in the bandit setting," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3250–3265, 2012.

[26] L. Li, W. Xu, T. Chen, G. B. Giannakis, and Q. Ling, "RSA: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 1544–1551, 2019.

[27] M. W. Seeger, C. K. Williams, and N. D. Lawrence, "Fast forward selection to speed up sparse Gaussian process regression," in *International Workshop on Artificial Intelligence and Statistics*, pp. 254–261, 2003.

[28] P. Rigollet and J.-C. Hütter, *High dimensional statistics*. Lecture notes for course, 2015.

[29] Z. Yuan and M. Zhu, "Lightweight distributed Gaussian process regression for online machine learning," *arXiv preprint arXiv:2105.04738*, 2021.

[30] T. Tao, *Topics in random matrix theory*, vol. 132. American Mathematical Society, 2012.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We enumerate the contributions in the introduction.
   (b) Did you describe the limitations of your work? [Yes] We discuss the future work in the conclusion (Section 5).
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We give all required assumptions in Sections 2.2 and 3.4.
   (b) Did you include complete proofs of all theoretical results? [Yes] We provide complete proofs of each lemma and theorem in Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We provide the simulation details in Section 4. We also include the code in the supplementary document.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We provide the training details including data partition in Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We run experiments multiple times and give error bars in Section 4.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We include the type of CPU in Section 4.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We use two real-world datasets, SARCOS and *Kin40k*, in Section 4, and we cite the creators.
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] Our work is not associated with research on human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Our work is not associated with research on human subjects.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Our work is not associated with research on human subjects.

A Proofs

In this section, we provide complete proofs for each lemma and theorem.

A.1 Proof of Lemma 1

Notice that $|\mathcal{I}(t)| = |\mathcal{M}(t) \bigcap \mathcal{C}(t)| \leq |\mathcal{M}(t)|$. By definition of $\mathcal{M}(t)$ in (6), we have $\mathcal{M}(t) = \mathcal{V} \setminus (\mathcal{T}^{\mu}_{\max}(t) \bigcup \mathcal{T}^{\mu}_{\min}(t))$. Definitions of $\mathcal{T}^{\mu}_{\max}(t)$ and $\mathcal{T}^{\mu}_{\min}(t)$ give $|\mathcal{T}^{\mu}_{\max}(t)| = |\mathcal{T}^{\mu}_{\min}(t)| = \lfloor \beta n \rfloor$. Then we have $|\mathcal{M}(t)| = |\mathcal{V} \setminus (\mathcal{T}^{\mu}_{\max}(t) \bigcup \mathcal{T}^{\mu}_{\min}(t))| = n - 2\lfloor \beta n \rfloor$, which indicates that $|\mathcal{I}(t)| \leq n - 2\lfloor \beta n \rfloor$.

Moreover, at time t, since $\mathcal{M}(t) \bigcup \mathcal{C}(t) \subseteq \mathcal{V}$, then it holds that $|\mathcal{M}(t) \bigcup \mathcal{C}(t)| \leq n$. By definition of $\mathcal{C}(t)$ in (6), we have $|\mathcal{C}(t)| = n - 2\lfloor \beta n \rfloor$. Therefore, we have $|\mathcal{I}(t)| = |\mathcal{M}(t) \bigcap \mathcal{C}(t)| = |\mathcal{M}(t)| + |\mathcal{C}(t)| - |\mathcal{M}(t) \bigcup \mathcal{C}(t)| \geq n - 4\lfloor \beta n \rfloor$. Since $\beta < \frac{1}{4}$, we have $|\mathcal{I}(t)| > 0$, i.e., $\mathcal{I}(t) \neq \emptyset$.

A.2 Proof of Lemma 2

Notice that $\mathcal{I}(t) = \mathcal{M}(t) \bigcap \mathcal{C}(t)$. Recall that in Step 2 of Section 3.2, $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ is sorted in ascending order, i.e., $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} < \check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i+1]}(t)}$ for $i = 1, 2, \ldots, n-1$. Recall that $\mathcal{T}^{\mu}_{\max}(t) = \{n - \lfloor \beta n \rfloor + 1, \ldots, n\}$. For any $q \in \mathcal{T}^{\mu}_{\max}(t)$ and $q' \in \mathcal{M}(t)$, we have $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[q]}(t)} > \check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[q']}(t)}$. Suppose that there exists $i \in \mathcal{M}(t)$ such that $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} > \max_{j \in \mathcal{V} \setminus \mathcal{B}} \left\{ \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)} \right\}$, then we have $i \in \mathcal{B}$. For all $q \in \mathcal{T}^{\mu}_{\max}(t)$, we have $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[q]}(t)} > \check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} > \max_{j \in \mathcal{V} \setminus \mathcal{B}} \left\{ \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)} \right\}$. Since $|\mathcal{T}^{\mu}_{\max}(t)| = \lfloor \beta n \rfloor$, then we have $\lfloor \alpha n \rfloor \geq \lfloor \beta n \rfloor + 1$. It contradicts with $\alpha \leq \beta$. Therefore, we have $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \leq \max_{j \in \mathcal{V} \setminus \mathcal{B}} \left\{ \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)} \right\}$ for all $i \in \mathcal{M}(t)$. Likewise, we have $\min_{j \in \mathcal{V} \setminus \mathcal{B}} \left\{ \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)} \right\} \leq \check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ for all $i \in \mathcal{M}(t)$.

Analogous to the proof of $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$, we conclude that $\min_{j \in \mathcal{V} \setminus \mathcal{B}} \left\{ \check{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)} \right\} \leq \check{\sigma}'^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \leq \max_{j \in \mathcal{V} \setminus \mathcal{B}} \left\{ \check{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \right\}$ for all $i \in \mathcal{C}(t)$.

A.3 Proof of Theorem 1

Part I: Roadmap of the proof: We first show in Lemma 3 that at time instant t, the local predictive mean of agent $i \in \mathcal{V}$ in the attack-free scenario is a sub-Gaussian random variable. Then notice that by triangular inequality, the prediction errors under attacks can be bounded by the magnitude of attacks plus the prediction errors in the attack-free case. Therefore, for $i \in \mathcal{I}(t)$, Lemma 4 uses concentration inequalities of sub-Gaussian random variables to quantify the upper bound of the

Byzantine attacks. We derive the upper bound of the prediction error in the attack-free case in Lemma 5.

Lemma 3 *Let Assumptions 2 and 3 hold. For agent $i \in \mathcal{V}$ and $\boldsymbol{z}_* \in \mathcal{Z}_*$, it holds that $\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ is a sub-Gaussian random variable.*

Proof: Monotonicity of Assumption 2 implies that $k(\boldsymbol{z}_*^{[i]}(t), \boldsymbol{z}_*^{[i]}(t)) = \kappa(0) = \sigma_f^2$. By Assumption 3, the prior mean is $\mu(\boldsymbol{z}_*) = \mu(\boldsymbol{z}_*^{[i]}(t)) = 0$. Hence, given observation $y^{[i]}_{\boldsymbol{z}_*^{[i]}(t)} = \eta(\boldsymbol{z}_*^{[i]}(t)) + e^{[i]}(t)$, by (5), the local predictive mean is computed as

$$\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} = \frac{\kappa(D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*})}{\sigma_f^2 + (\sigma_e^{[i]})^2} y^{[i]}_{\boldsymbol{z}_*^{[i]}(t)} = \frac{\kappa(D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*})}{\sigma_f^2 + (\sigma_e^{[i]})^2} \left(\eta(\boldsymbol{z}_*^{[i]}(t)) + e^{[i]}(t)\right). \tag{9}$$

For agent $i \in \mathcal{V}$, we denote the expectation and variance of $\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ by $\mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]$ and $Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]$, respectively. Notice that, in (9), $e^{[i]}(t)$ is the only random variable and $e^{[i]}(t) \sim \mathcal{N}(0, (\sigma_e^{[i]})^2)$, this implies $\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \sim \mathcal{N}\left(\frac{\kappa(D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*})}{\sigma_f^2+(\sigma_e^{[i]})^2}\eta(\boldsymbol{z}_*^{[i]}(t)), \left(\frac{\kappa(D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*})}{\sigma_f^2+(\sigma_e^{[i]})^2}\right)^2 (\sigma_e^{[i]})^2\right)$. Then for $\lambda_i \in \mathbb{R}$, we conduct the following algebraic calculations

$$\begin{aligned}
&\mathbb{E}\left[\exp(\lambda_i(\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]))\right] \\
&= \int_{-\infty}^{+\infty} \frac{\exp\left(-\frac{\left(\mu - \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]\right)^2}{2Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}\right)}{\sqrt{2\pi Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}} \exp\left(\lambda_i\left(\mu - \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]\right)\right) d\mu \\
&= \frac{\exp\left(-\lambda_i \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]\right)}{\sqrt{2\pi Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}} \int_{-\infty}^{+\infty} \exp\left(\lambda_i \mu - \frac{\left(\mu - \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]\right)^2}{2Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}\right) d\mu.
\end{aligned}$$

Note that

$$\begin{aligned}
&\exp\left(-\lambda_i \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]\right) \exp\left(\lambda_i \mu - \frac{\left(\mu - \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]\right)^2}{2Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}\right) \\
&= \exp\left(-\lambda_i\left(\mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right] - \mu\right) - \frac{\left(\mu - \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]\right)^2}{2Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}\right) \\
&= \exp\left(\frac{-\left(2\lambda_i \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right] - 2\lambda_i \mu\right) Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}{2Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}\right) \\
&\quad \times \exp\left(-\frac{\mu^2 - 2\mu \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right] + \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]^2}{2Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}\right) \\
&= \exp\left(\frac{\left(\mu - \left(\lambda_i Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right] + \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]\right)\right)^2}{-2Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}\right) \\
&\quad \times \exp\left(-\lambda_i \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right] - \frac{\mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]^2}{2Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}\right) \exp\left(\frac{\left(\lambda_i Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right] + \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]\right)^2}{2Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}\right) \\
&= \exp\left(\frac{\lambda_i^2 Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}{2}\right) \exp\left(\frac{\left(\mu - \left(\lambda_i Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right] + \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]\right)\right)^2}{-2Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}\right).
\end{aligned}$$

Then we have

$$\mathbb{E}\left[\exp\left(\lambda_i(\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right])\right)\right]$$

$$= \exp\left(\frac{\lambda_i^2 Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}{2}\right) \underbrace{\frac{\int_{-\infty}^{+\infty} \exp\left(\frac{\left(\mu-\left(\lambda_i Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]+\mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]\right)\right)^2}{-2Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}\right) d\mu}{\sqrt{2\pi Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}}}_{=1}$$

$$= \exp\left(\frac{\lambda_i^2 Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}{2}\right). \tag{10}$$

The term $\frac{\exp\left(\frac{\left(\mu-\left(\lambda_i Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]+\mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]\right)\right)^2}{-2Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}\right)}{\sqrt{2\pi Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]}}$ is a Gaussian probability density function with mean $\lambda_i Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right] + \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]$ and variance $Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right]$. Recall that $\sigma^2 = \left(\frac{\sigma_f^2 \sigma_e^{\max}}{\sigma_f^2+(\sigma_e^{\min})^2}\right)^2$. By Assumption 2, it holds that the kernel function $\kappa(\cdot)$ is monotonically decreasing and $\kappa(0) = \sigma_f^2$. Therefore, we have $Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right] = \left(\frac{\kappa(D_{\boldsymbol{z}_*}^{\mathcal{Z}^{[i]}(t)})}{\sigma_f^2+(\sigma_e^{[i]})^2}\sigma_e^{[i]}\right)^2 \leq \left(\frac{\sigma_f^2\sigma_e^{\max}}{\sigma_f^2+(\sigma_e^{\min})^2}\right)^2 = \sigma^2$. Substituting $Var\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right] \leq \sigma^2$ into (10) yields $\mathbb{E}\left[\exp\left(\lambda_i(\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \mathbb{E}\left[\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right])\right)\right] \leq \exp\left(\frac{\lambda_i^2\sigma^2}{2}\right)$. Thus by Definition 1, we conclude that for agent $i \in \mathcal{V}$, it holds that $\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ is a sub-Gaussian random variable. ■

Lemma 4 *Let $0 < \alpha \leq \beta < \frac{1}{4}$ and Assumption 2 hold. For all $\boldsymbol{z}_* \in \mathcal{Z}_*$ and $0 < \delta < 1$, with probability at least $1-\delta$, it holds that* $\frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \left|\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right| \leq \frac{2\alpha(\sqrt{2\sigma^2(\ln(2n)-\ln\delta)}+\frac{\sigma_f^2\|\eta\|_\infty}{\sigma_f^2+(\sigma_e^{\min})^2})}{1-4\beta}\frac{\sigma_f^4+\sigma_f^2(\sigma_e^{\max})^2-\kappa(d^{\max}(t))^2}{\sigma_f^2(\sigma_e^{\min})^2}$.

Proof: We denote by $\mathcal{F}(t) \triangleq \mathcal{I}(t)\bigcap(\mathcal{V}\backslash\mathcal{B})$ the set of benign agents in the set $\mathcal{I}(t)$. That is, $\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} = \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ for all $i \in \mathcal{F}(t)$, therefore it holds that $\frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{F}(t)} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \left|\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right| = 0$. Recall that $\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)} = \frac{|\mathcal{I}(t)|}{\sum_{j\in\mathcal{I}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)}}$. We have

$$\begin{aligned}
&\frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \left|\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right| \\
&= \frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)\bigcap\mathcal{B}} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \left|\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right| \\
&\quad + \underbrace{\frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{F}(t)} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \left|\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right|}_{=0} \\
&= \frac{\sum_{i\in\mathcal{I}(t)\bigcap\mathcal{B}} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \left|\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right|}{\sum_{j\in\mathcal{I}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)}} \\
&\leq \frac{\sum_{i\in\mathcal{I}(t)\bigcap\mathcal{B}} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \left(\left|\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right| + \left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right|\right)}{\sum_{j\in\mathcal{I}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)}}.
\end{aligned} \tag{11}$$

In the remaining proof, we find the upper bound of $\left|\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right| + \left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right|$, and characterize the lower and upper bounds of $\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$.

1) *The upper bound of* $\left|\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right| + \left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right|$. By Lemma 2, we have $\left|\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right| \leq \max_{i\in\mathcal{V}\backslash\mathcal{B}}\left\{\left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right|\right\}$ for all $i \in \mathcal{I}(t)$. Then $\left|\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right| + \left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right| \leq 2\max_{i\in\mathcal{V}\backslash\mathcal{B}}\left\{\left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right|\right\}$ for all $i \in \mathcal{I}(t)$. By Lemma 3, for all $i \in \mathcal{V}$, $\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ is a sub-Gausian random variable. Since $|\mathcal{V}\backslash\mathcal{B}| = n - \lfloor\alpha n\rfloor$, by maximal inequality (Theorem 1.14 on page 25 in [28]), for any $\epsilon_1 > 0$, we have

$$P\left\{\max_{i\in\mathcal{V}\backslash\mathcal{B}}\left\{\left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \frac{\kappa(D_{\boldsymbol{z}_*}^{\mathcal{Z}^{[i]}(t)})}{\sigma_f^2 + (\sigma_e^{[i]})^2}\eta(\boldsymbol{z}_*^{[i]}(t))\right|\right\} \geq \epsilon_1\right\} \leq 2(n - \lfloor\alpha n\rfloor)e^{-\frac{\epsilon_1^2}{2\sigma^2}} \leq 2ne^{-\frac{\epsilon_1^2}{2\sigma^2}}.$$

For $0 < \delta < 1$, choosing $\epsilon_1 \triangleq \sqrt{2\sigma^2(\ln(2n) - \ln\delta)}$, with probability at least $1-\delta$, we have

$$\max_{i\in\mathcal{V}\backslash\mathcal{B}}\left\{\left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \frac{\kappa(D_{\boldsymbol{z}_*}^{\mathcal{Z}^{[i]}(t)})}{\sigma_f^2 + (\sigma_e^{[i]})^2}\eta(\boldsymbol{z}_*^{[i]}(t))\right|\right\} \leq \sqrt{2\sigma^2(\ln(2n) - \ln\delta)}.$$

Since triangular inequality renders

$$\left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right| \leq \left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \frac{\kappa(D_{\boldsymbol{z}_*}^{\mathcal{Z}^{[i]}(t)})}{\sigma_f^2 + (\sigma_e^{[i]})^2}\eta(\boldsymbol{z}_*^{[i]}(t))\right| + \left|\frac{\kappa(D_{\boldsymbol{z}_*}^{\mathcal{Z}^{[i]}(t)})}{\sigma_f^2 + (\sigma_e^{[i]})^2}\eta(\boldsymbol{z}_*^{[i]}(t))\right|,$$

we have

$$\begin{aligned}\max_{i\in\mathcal{V}\backslash\mathcal{B}}\left\{\left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right|\right\} \leq& \max_{i\in\mathcal{V}\backslash\mathcal{B}}\left\{\left|\frac{\kappa(D_{\boldsymbol{z}_*}^{\mathcal{Z}^{[i]}(t)})}{\sigma_f^2 + (\sigma_e^{[i]})^2}\eta(\boldsymbol{z}_*^{[i]}(t))\right|\right\} \\ &+ \max_{i\in\mathcal{V}\backslash\mathcal{B}}\left\{\left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \frac{\kappa(D_{\boldsymbol{z}_*}^{\mathcal{Z}^{[i]}(t)})}{\sigma_f^2 + (\sigma_e^{[i]})^2}\eta(\boldsymbol{z}_*^{[i]}(t))\right|\right\},\end{aligned}$$

which implies that with probability at least $1-\delta$,

$$\max_{i\in\mathcal{V}\backslash\mathcal{B}}\left\{\left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right|\right\} \leq \sqrt{2\sigma^2(\ln(2n) - \ln\delta)} + \max_{i\in\mathcal{V}\backslash\mathcal{B}}\left\{\left|\frac{\kappa(D_{\boldsymbol{z}_*}^{\mathcal{Z}^{[i]}(t)})}{\sigma_f^2 + (\sigma_e^{[i]})^2}\eta(\boldsymbol{z}_*^{[i]}(t))\right|\right\}.$$

Since $|\eta(\boldsymbol{z}_*^{[i]}(t))| \leq \|\eta\|_\infty$, by monotonicity of $\kappa(\cdot)$ in Assumption 2, it holds that $\max_{i\in\mathcal{V}\backslash\mathcal{B}}\left\{\left|\frac{\kappa(D_{\boldsymbol{z}_*}^{\mathcal{Z}^{[i]}(t)})}{\sigma_f^2+(\sigma_e^{[i]})^2}\eta(\boldsymbol{z}_*^{[i]}(t))\right|\right\} \leq \frac{\sigma_f^2\|\eta\|_\infty}{\sigma_f^2+(\sigma_e^{\min})^2}$. Then with probability at least $1-\delta$, we have

$$\max_{i\in\mathcal{V}\backslash\mathcal{B}}\left\{\left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right|\right\} \leq \sqrt{2\sigma^2(\ln(2n) - \ln\delta)} + \frac{\sigma_f^2\|\eta\|_\infty}{\sigma_f^2 + (\sigma_e^{\min})^2}.$$

Therefore, with probability at least $1-\delta$, we have

$$\begin{aligned}\left|\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right| + \left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right| &\leq 2\max_{i\in\mathcal{V}\backslash\mathcal{B}}\left\{\left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right|\right\} \\ &\leq 2(\sqrt{2\sigma^2(\ln(2n) - \ln\delta)} + \frac{\sigma_f^2\|\eta\|_\infty}{\sigma_f^2 + (\sigma_e^{\min})^2}). \qquad (12)\end{aligned}$$

2) *The lower and upper bounds of* $\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$. Lemma 2 renders that $\min_{j\in\mathcal{V}\backslash\mathcal{B}}\left\{\check{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)}\right\} \leq \check{\sigma}'^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \leq \max_{j\in\mathcal{V}\backslash\mathcal{B}}\left\{\check{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)}\right\}$ for all $i \in \mathcal{I}(t)$, then we have $\left(\max_{j\in\mathcal{V}\backslash\mathcal{B}}\left\{\check{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)}\right\}\right)^{-1} \leq \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \leq \left(\min_{j\in\mathcal{V}\backslash\mathcal{B}}\left\{\check{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)}\right\}\right)^{-1}$ for all $i \in \mathcal{I}(t)$. Theorem IV.3 in [29] gives $\frac{\sigma_f^2(\sigma_e^{[i]})^2}{\sigma_f^2+(\sigma_e^{[i]})^2} \leq \check{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \leq \sigma_f^2 - \frac{\kappa(d^{[i]}(t))^2}{\sigma_f^2+(\sigma_e^{[i]})^2}$ for all $\boldsymbol{z}_* \in \mathcal{Z}_*$. By monotonicity of $\kappa(\cdot)$ in Assumption 2, it holds that $\frac{\sigma_f^2(\sigma_e^{\min})^2}{\sigma_f^2+(\sigma_e^{\max})^2} \leq \check{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \leq \sigma_f^2 - \frac{\kappa(d^{\max}(t))^2}{\sigma_f^2+(\sigma_e^{\max})^2}$, which implies that for any $i \in \mathcal{I}(t), \boldsymbol{z}_* \in \mathcal{Z}_*$,

$$\frac{\sigma_f^2 + (\sigma_e^{\max})^2}{\sigma_f^4 + \sigma_f^2(\sigma_e^{\max})^2 - \kappa(d^{\max}(t))^2} \leq \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \leq \frac{\sigma_f^2 + (\sigma_e^{\max})^2}{\sigma_f^2(\sigma_e^{\min})^2}.$$

Lemma 1 shows that $|\mathcal{I}(t)| \geq n - 4\lfloor \beta n \rfloor$. Since $\lfloor \beta n \rfloor \leq \beta n$, it indicates that $|\mathcal{I}(t)| \geq (1-4\beta)n$. Then we have

$$\sum_{j \in \mathcal{I}(t)} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[j]}(t)} \geq \frac{(1-4\beta)n\left(\sigma_f^2 + (\sigma_e^{\max})^2\right)}{\sigma_f^4 + \sigma_f^2(\sigma_e^{\max})^2 - \kappa(d^{\max}(t))^2}. \tag{13}$$

Since $|\mathcal{I}(t) \bigcap \mathcal{B}| \leq |\mathcal{B}| = \lfloor \alpha n \rfloor \leq \alpha n$, by (12), with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sum_{i \in \mathcal{I}(t) \bigcap \mathcal{B}} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \left(\left| \check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \right| + \left| \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \right| \right) \\
&\leq 2\alpha n \left(\sqrt{2\sigma^2(\ln(2n) - \ln\delta)} + \frac{\sigma_f^2 \|\eta\|_\infty}{\sigma_f^2 + (\sigma_e^{\min})^2}\right) \frac{\sigma_f^2 + (\sigma_e^{\max})^2}{\sigma_f^2(\sigma_e^{\min})^2}.
\end{aligned} \tag{14}$$

Combining (13) and (14) with (11) renders that with probability at least $1-\delta$, it holds that

$$\begin{aligned}
&\frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|} \sum_{i \in \mathcal{I}(t)} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \left| \check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \right| \\
&\leq \frac{2\alpha\left(\sqrt{2\sigma^2(\ln(2n) - \ln\delta)} + \frac{\sigma_f^2\|\eta\|_\infty}{\sigma_f^2 + (\sigma_e^{\min})^2}\right)}{1-4\beta} \frac{\sigma_f^4 + \sigma_f^2(\sigma_e^{\max})^2 - \kappa(d^{\max}(t))^2}{\sigma_f^2(\sigma_e^{\min})^2}.
\end{aligned}$$

■

The following Lemma characterizes the upper bound of the prediction error in the attack-free case.

Lemma 5 *Suppose Assumptions 2 and 3 hold. For $\boldsymbol{z}_* \in \mathcal{Z}_*$, with probability at least $1-\delta$, it holds that $\frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|} \sum_{i \in \mathcal{I}(t)} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \left| \check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \eta(\boldsymbol{z}_*) \right| \leq (1 - \frac{\kappa(d^{\max}(t))}{\sigma_f^2 + (\sigma_e^{\max})^2})\|\eta\|_\infty + \frac{\sigma_f^2}{\sigma_f^2 + (\sigma_e^{\min})^2}\ell_\eta d^{\max}(t) + \sqrt{2\sigma^2(\ln 2 - \ln\delta)}$.*

Proof: Recall that $\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} = \frac{\kappa(D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*})}{\sigma_f^2 + (\sigma_e^{[i]})^2} y^{[i]}_{\boldsymbol{z}^{[i]}_*(t)}$. Then for $i \in \mathcal{V}$, we have

$$\begin{aligned}
\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \eta(\boldsymbol{z}_*) &= (1 - \frac{\kappa(D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*})}{\sigma_f^2 + (\sigma_e^{[i]})^2})(-\eta(\boldsymbol{z}_*)) + \frac{\kappa(D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*})}{\sigma_f^2 + (\sigma_e^{[i]})^2}(y^{[i]}_{\boldsymbol{z}^{[i]}_*(t)} - \eta(\boldsymbol{z}^{[i]}_*(t))) \\
&\quad + \frac{\kappa(D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*})}{\sigma_f^2 + (\sigma_e^{[i]})^2}(\eta(\boldsymbol{z}^{[i]}_*(t)) - \eta(\boldsymbol{z}_*)).
\end{aligned}$$

By triangular inequality, we have

$$\begin{aligned}
|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \eta(\boldsymbol{z}_*)| &\leq \frac{\kappa(D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*})}{\sigma_f^2 + (\sigma_e^{[i]})^2}|\eta(\boldsymbol{z}^{[i]}_*(t)) - \eta(\boldsymbol{z}_*)| + \frac{\kappa(D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*})}{\sigma_f^2 + (\sigma_e^{[i]})^2}|y^{[i]}_{\boldsymbol{z}^{[i]}_*(t)} - \eta(\boldsymbol{z}^{[i]}_*(t))| \\
&\quad + (1 - \frac{\kappa(D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*})}{\sigma_f^2 + (\sigma_e^{[i]})^2})|\eta(\boldsymbol{z}_*)|.
\end{aligned} \tag{15}$$

We analyze the upper bound of each term on the right-hand side of the inequality (15).

Term 1. Recall that $\boldsymbol{z}^{[i]}_*(t) \in$ proj$(\boldsymbol{z}_*, \mathcal{Z}^{[i]}(t))$. Using the Lipschitz continuity of η in Assumption 3 gives

$$\frac{\kappa(D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*})}{\sigma_f^2 + (\sigma_e^{[i]})^2} \left| \eta(\boldsymbol{z}^{[i]}_*(t)) - \eta(\boldsymbol{z}_*) \right| \leq \frac{\kappa(D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*})}{\sigma_f^2 + (\sigma_e^{[i]})^2}\ell_\eta D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*} \leq \frac{\sigma_f^2}{\sigma_f^2 + (\sigma_e^{\min})^2}\ell_\eta d^{\max}(t). \tag{16}$$

Term 2. Recall that $\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ follows a Gaussian probability distribution and $\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \sim \mathcal{N}\left(\frac{\kappa(D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*})}{\sigma_f^2 + (\sigma_e^{[i]})^2}\eta(\boldsymbol{z}^{[i]}_*(t)), \left(\frac{\kappa(D^{\mathcal{Z}^{[i]}(t)}_{\boldsymbol{z}_*})}{\sigma_f^2 + (\sigma_e^{[i]})^2}\right)^2 (\sigma_e^{[i]})^2\right)$. By Lemma 3, for all $i \in \mathcal{V}$, we have that

$\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ is a sub-Gaussian random variable. Then by concentration inequality of the sub-Gaussian random variable (see Lemma 1.3 of [28]), for any $\epsilon_2 > 0$, we have

$$P\left\{\left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \frac{\kappa(D_{\boldsymbol{z}_*}^{\mathcal{Z}^{[i]}(t)})}{\sigma_f^2 + (\sigma_e^{[i]})^2}\eta(\boldsymbol{z}_*^{[i]}(t))\right| > \epsilon_2\right\} \leq 2e^{-\frac{\epsilon_2^2}{2\sigma^2}}.$$

Combining the above inequality with $\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} = \frac{\kappa(D_{\boldsymbol{z}_*}^{\mathcal{Z}^{[i]}(t)})}{\sigma_f^2+(\sigma_e^{[i]})^2} y_{\boldsymbol{z}_*^{[i]}(t)}^{[i]}$, for $0 < \delta < 1$, choosing $\epsilon_2 \triangleq \sqrt{2\sigma^2(\ln 2 - \ln \delta)}$, with probability at least $1-\delta$, it holds

$$\frac{\kappa(D_{\boldsymbol{z}_*}^{\mathcal{Z}^{[i]}(t)})}{\sigma_f^2 + (\sigma_e^{[i]})^2}|y_{\boldsymbol{z}_*^{[i]}(t)}^{[i]} - \eta(\boldsymbol{z}_*^{[i]}(t))| = \left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \frac{\kappa(D_{\boldsymbol{z}_*}^{\mathcal{Z}^{[i]}(t)})}{\sigma_f^2 + (\sigma_e^{[i]})^2}\eta(\boldsymbol{z}_*^{[i]}(t))\right| \leq \sqrt{2\sigma^2(\ln 2 - \ln \delta)}. \tag{17}$$

Term 3. We have $|\eta(\boldsymbol{z}_*)| \leq \|\eta\|_\infty$. By monotonicity of $\kappa(\cdot)$ in Assumption 2, it gives

$$(1 - \frac{\kappa(D_{\boldsymbol{z}_*}^{\mathcal{Z}^{[i]}(t)})}{\sigma_f^2 + (\sigma_e^{[i]})^2})|\eta(\boldsymbol{z}_*)| \leq (1 - \frac{\kappa(d^{\max}(t))}{\sigma_f^2 + (\sigma_e^{\max})^2})\|\eta\|_\infty. \tag{18}$$

Therefore, applying the inequalities (16), (17) and (18) to (15), for $0 < \delta < 1$, with probability at least $1-\delta$, we have that for all $i \in \mathcal{I}(t)$,

$$\left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \eta(\boldsymbol{z}_*)\right| \leq (1 - \frac{\kappa(d^{\max}(t))}{\sigma_f^2 + (\sigma_e^{\max})^2})\|\eta\|_\infty + \frac{\sigma_f^2 \ell_\eta d^{\max}(t)}{\sigma_f^2 + (\sigma_e^{\min})^2} + \sqrt{2\sigma^2(\ln 2 - \ln \delta)}. \tag{19}$$

By (7), we have $0 < \frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} < 1$ and $\frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} = 1$, which implies that with probability at least $1-\delta$, it holds that $\frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \eta(\boldsymbol{z}_*)\right| \leq (1 - \frac{\kappa(d^{\max}(t))}{\sigma_f^2+(\sigma_e^{\max})^2})\|\eta\|_\infty + \frac{\sigma_f^2\ell_\eta d^{\max}(t)}{\sigma_f^2+(\sigma_e^{\min})^2} + \sqrt{2\sigma^2(\ln 2 - \ln\delta)}$. ■

With Lemmas 3, 4 and 5, we now proceed to complete the proof of part I in Theorem 1.

Proof of part I in Theorem 1: Note that, given (7), we have

$$\begin{aligned}
\left|\hat{\mu}_{\boldsymbol{z}_*|\mathcal{D}(t)} - \eta(\boldsymbol{z}_*)\right| &= \left|\frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)}\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \eta(\boldsymbol{z}_*)\right| \\
&= \left|\frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)}\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)}\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right. \\
&\quad \left. + \frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)}\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \eta(\boldsymbol{z}_*)\right|.
\end{aligned} \tag{20}$$

Since $\frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} = 1$, then this implies that $\frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)}\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \eta(\boldsymbol{z}_*) = \frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\left(\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \eta(\boldsymbol{z}_*)\right)$. Therefore, by triangular inequality, (20) is upper bounded as

$$\begin{aligned}
&\left|\frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)}\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \eta(\boldsymbol{z}_*)\right| \\
&\leq \frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\left|\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} - \eta(\boldsymbol{z}_*)\right|
\end{aligned}$$

$$+\frac{\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\left|\check{\mu}'_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}-\check{\mu}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right|.$$

Then, combining this with Lemmas 4 and 5, we complete the proof of part I.

Part II: We give the upper bound and lower bound of $\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}$ as follows:

1) Upper bound. Recall that $\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}=\frac{|\mathcal{I}(t)|}{\sum_{i\in\mathcal{I}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}}$. Note that $f(x)=\frac{1}{x}$ is a convex function for $x>0$. By Jensen's inequality (see page 21 in [30]), we have $f(\frac{1}{n}\sum_{i=1}^n x_i)\leq\frac{1}{n}\sum_{i=1}^n f(x_i)$. Then plugging in $x_i=\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$, we have

$$\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}=\frac{|\mathcal{I}(t)|}{\sum_{i\in\mathcal{I}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}}=f(\frac{\sum_{i\in\mathcal{I}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}}{|\mathcal{I}(t)|})\leq\frac{1}{|\mathcal{I}(t)|}\sum_{i\in\mathcal{I}(t)}\check{\sigma}'^{2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}.$$

It suffices to show the upper bound of $\sum_{i\in\mathcal{I}(t)}\check{\sigma}'^{2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$. We decompose the agent set $\mathcal{I}(t)$ into two subsets $\mathcal{F}(t)$ and $\mathcal{I}(t)\bigcap\mathcal{B}$ where $\mathcal{F}(t)\triangleq\mathcal{I}(t)\bigcap(\mathcal{V}\backslash\mathcal{B})$ contains the benign agents and $\mathcal{I}(t)\bigcap\mathcal{B}$ contains the Byzantine agents, then we have

$$\sum_{i\in\mathcal{I}(t)}\check{\sigma}'^{2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}=\sum_{i\in\mathcal{F}(t)}\check{\sigma}'^{2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}+\sum_{i\in\mathcal{I}(t)\bigcap\mathcal{B}}\check{\sigma}'^{2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}. \tag{21}$$

We proceed to analyze each term on the right-hand side of (21).

First, notice that $\mathcal{F}(t)$ is the set of benign agents, hence it holds that $\check{\sigma}'^{2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}=\check{\sigma}^{2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ for all $i\in\mathcal{F}(t)$. According to Theorem IV.3 in [29], we have $\check{\sigma}^{2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\leq\sigma_f^2-\frac{\kappa(d^{[i]}(t))^2}{\sigma_f^2+(\sigma_e^{[i]})^2}$ for all $\boldsymbol{z}_*\in\mathcal{Z}_*$. Monotonicity of $\kappa(\cdot)$ in Assumption 2 shows that $\kappa(d^{\max}(t))^2\leq\kappa(d^{[i]}(t))^2$, which indicates $\frac{\kappa(d^{\max}(t))^2}{\sigma_f^2+(\sigma_e^{\max})^2}\leq\frac{\kappa(d^{[i]}(t))^2}{\sigma_f^2+(\sigma_e^{[i]})^2}$. Therefore, we have

$$\sum_{i\in\mathcal{F}(t)}\check{\sigma}'^{2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\leq\sum_{i\in\mathcal{F}(t)}\left(\sigma_f^2-\frac{\kappa(d^{[i]}(t))^2}{\sigma_f^2+(\sigma_e^{[i]})^2}\right)\leq|\mathcal{F}(t)|\left(\sigma_f^2-\frac{\kappa(d^{\max}(t))^2}{\sigma_f^2+(\sigma_e^{\max})^2}\right). \tag{22}$$

Second, by Lemma 2, we have

$$\begin{aligned}\sum_{i\in\mathcal{I}(t)\bigcap\mathcal{B}}\check{\sigma}'^{2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}&\leq\sum_{i\in\mathcal{I}(t)\bigcap\mathcal{B}}\max_{i\in\mathcal{V}\backslash\mathcal{B}}\left\{\check{\sigma}^{2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right\}\\&\leq\sum_{i\in\mathcal{I}(t)\bigcap\mathcal{B}}\max_{i\in\mathcal{V}\backslash\mathcal{B}}\left\{\sigma_f^2-\frac{\kappa(d^{[i]}(t))^2}{\sigma_f^2+(\sigma_e^{[i]})^2}\right\}=|\mathcal{I}(t)\bigcap\mathcal{B}|\left(\sigma_f^2-\frac{\kappa(d^{\max}(t))^2}{\sigma_f^2+(\sigma_e^{\max})^2}\right).\end{aligned} \tag{23}$$

Therefore, combining (21) with (22) and (23), the upper bound of $\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}$ is given as

$$\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)}\leq\sigma_f^2-\frac{\kappa(d^{\max}(t))^2}{\sigma_f^2+(\sigma_e^{\max})^2}. \tag{24}$$

2) Lower bound. Similar to (21), we have

$$\sum_{i\in\mathcal{I}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}=\sum_{i\in\mathcal{F}(t)}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}+\sum_{i\in\mathcal{I}(t)\bigcap\mathcal{B}}\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}. \tag{25}$$

First, it holds that $\check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}=\check{\sigma}^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$ for all $i\in\mathcal{F}(t)$. Theorem IV.3 in [29] gives $\check{\sigma}^{2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\geq\frac{\sigma_f^2(\sigma_e^{[i]})^2}{\sigma_f^2+(\sigma_e^{[i]})^2}$ for all $\boldsymbol{z}_*\in\mathcal{Z}_*$. Then following the logic of deriving the above upper bound, we have

$$\sum_{i\in\mathcal{F}(t)}\check{\sigma}^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\leq\sum_{i\in\mathcal{F}(t)}\frac{\sigma_f^2+(\sigma_e^{[i]})^2}{\sigma_f^2(\sigma_e^{[i]})^2}\leq|\mathcal{F}(t)|\frac{\sigma_f^2+(\sigma_e^{\max})^2}{\sigma_f^2(\sigma_e^{\min})^2}. \tag{26}$$

Second, Lemma 2 implies the following inequality

$$\sum_{i\in\mathcal{I}(t)\bigcap\mathcal{B}} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \leq \sum_{i\in\mathcal{I}(t)\bigcap\mathcal{B}} \left(\min_{i\in\mathcal{V}\backslash\mathcal{B}} \left\{\check{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}\right\}\right)^{-1} \leq \sum_{i\in\mathcal{I}(t)\bigcap\mathcal{B}} \left(\min_{i\in\mathcal{V}\backslash\mathcal{B}} \left\{\frac{\sigma_f^2(\sigma_e^{[i]})^2}{\sigma_f^2+(\sigma_e^{[i]})^2}\right\}\right)^{-1}$$
$$\leq |\mathcal{I}(t)\bigcap\mathcal{B}|\frac{\sigma_f^2+(\sigma_e^{\max})^2}{\sigma_f^2(\sigma_e^{\min})^2}. \tag{27}$$

Therefore, combining (25) with (26) and (27) yields an upper bound of $\sum_{i\in\mathcal{I}(t)} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}$, i.e., $\sum_{i\in\mathcal{I}(t)} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)} \leq |\mathcal{I}(t)|\frac{\sigma_f^2+(\sigma_e^{\max})^2}{\sigma_f^2(\sigma_e^{\min})^2}$. Recall that $\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)} = \frac{|\mathcal{I}(t)|}{\sum_{i\in\mathcal{I}(t)} \check{\sigma}'^{-2}_{\boldsymbol{z}_*|\mathcal{D}^{[i]}(t)}}$, then we have

$$\hat{\sigma}^2_{\boldsymbol{z}_*|\mathcal{D}(t)} \geq \frac{\sigma_f^2(\sigma_e^{\min})^2}{\sigma_f^2+(\sigma_e^{\max})^2}. \tag{28}$$

Thus, combining (24) and (28), the proof is complete.