Generalized Robust Bayesian Committee Machine for Large-scale Gaussian Process Regression

Haitao Liu \textsuperscript{1} Jianfei Cai \textsuperscript{2} Yi Wang \textsuperscript{3} Yew-Soon Ong \textsuperscript{2,4}

Abstract

In order to scale standard Gaussian process (GP) regression to large-scale datasets, aggregation models employ factorized training process and then combine predictions from distributed experts. The state-of-the-art aggregation models, however, either provide inconsistent predictions or require time-consuming aggregation process. We first prove the inconsistency of typical aggregations using disjoint or random data partition, and then present a consistent yet efficient aggregation model for large-scale GP. The proposed model inherits the advantages of aggregations, e.g., closed-form inference and aggregation, parallelization and distributed computing. Furthermore, theoretical and empirical analyses reveal that the new aggregation model performs better due to the consistent predictions that converge to the true underlying function when the training size approaches infinity.

1. Introduction

Gaussian process (GP) (Rasmussen & Williams, 2006) is a well-known statistical learning model extensively used in various scenarios, e.g., regression, classification, optimization (Shahriari et al., 2016), visualization (Lawrence, 2005), active learning (Fu et al., 2013; Liu et al., 2017) and multi-task learning (Alvarez et al., 2012; Liu et al., 2018). Given the training set $X = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ and the observation set $y = \{y(x_i) \in \mathbb{R}\}_{i=1}^n$, as an approximation of the underlying function $\eta : \mathbb{R}^d \rightarrow \mathbb{R}$, GP provides informative predictive distributions at test points.

However, the most prominent weakness of the full GP is that it scales poorly with the training size. Given $n$ data points, the time complexity of a standard GP paradigm scales as $O(n^3)$ in the training process due to the inversion of an $n \times n$ covariance matrix; it scales as $O(n^2)$ in the prediction process due to the matrix-vector operation. This weakness confines the full GP to training data of size $O(10^4)$.

To cope with large-scale regression, various computationally efficient approximations have been presented. The sparse approximations reviewed in (Quiñonero-Candela & Rasmussen, 2005) employ $m (m \ll n)$ inducing points to summarize the whole training data (Seeger et al., 2003; Snelson & Ghahramani, 2006; 2007; Titsias, 2009; Bauer et al., 2016), thus reducing the training complexity of full GP to $O(n m^2)$ and the predicting complexity to $O(n m)$. The complexity can be further reduced through distributed inference, stochastic variational inference or Kronecker structure (Hensman et al., 2013; Gal et al., 2014; Wilson & Nickisch, 2015; Hoang et al., 2016; Peng et al., 2017). A main drawback of sparse approximations, however, is that the representational capability is limited by the number of inducing points (Moore & Russell, 2015). For example, for a quick-varying function, the sparse approximations need many inducing points to capture the local structures. That is, this kind of scheme has not reduced the scaling of the complexity (Bui & Turner, 2014).

The method exploited in this article belongs to the aggregation models (Hinton, 2002; Tresp, 2000; Cao & Fleet, 2014; Deisenroth & Ng, 2015; Rullière et al., 2017), also known as consensus statistical methods (Genest & Zidek, 1986; Ranjan & Gneiting, 2010). This kind of scheme produces the final predictions by the aggregation of $M$ submodels (GP experts) respectively trained on the subsets $\{D_i = \{X_i, y_i\}\}_{i=1}^M$ of $D = \{X, y\}$, thus distributing the computations to “local” experts. Particularly, due to the product of experts, the aggregation scheme derives a factorized marginal likelihood for efficient training; and then it combines the experts’ posterior distributions according to a certain aggregation criterion. In comparison to sparse approximations, the aggregation models (i) operate directly on the full training data, (ii) require no additional

\footnotesize
\textsuperscript{1}Rolls-Royce@NTU Corporate Lab, Nanyang Technological University, Singapore 637460 \textsuperscript{2}School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 \textsuperscript{3}Applied Technology Group, Rolls-Royce Singapore, 6 Seletar Aerospace Rise, Singapore 797575 \textsuperscript{4}Data Science and Artificial Intelligence Research Center, Nanyang Technological University, Singapore 639798. Correspondence to: Haitao Liu <hliu@ntu.edu.sg>.

Proceedings of the 35\textsuperscript{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

$\mathbf{X} = \{x_i \in \mathbb{R}^d\}_{i=1}^n$ and the observa-
$\mathbf{y} = \{y(x_i) \in \mathbb{R}\}_{i=1}^n$, as an approximation of the
underlying function $\eta : \mathbb{R}^d \rightarrow \mathbb{R}$, GP provides informative
predictive distributions at test points.

However, the most prominent weakness of the full GP is that
it scales poorly with the training size. Given $n$ data points,
the time complexity of a standard GP paradigm scales as
$O(n^3)$ in the training process due to the inversion of an
$n \times n$ covariance matrix; it scales as $O(n^2)$ in the prediction
process due to the matrix-vector operation. This weakness
confines the full GP to training data of size $O(10^4)$.

To cope with large-scale regression, various computa-
tionally efficient approximations have been presented. The
sparse approximations reviewed in (Quiñonero-Candela &
Rasmussen, 2005) employ $m (m \ll n)$ inducing points to
summarize the whole training data (Seeger et al., 2003; Snel-
son & Ghahramani, 2006; 2007; Titsias, 2009; Bauer et al.,
2016), thus reducing the training complexity of full GP to
$O(n m^2)$ and the predicting complexity to $O(n m)$. The
complexity can be further reduced through distributed infer-
ence, stochastic variational inference or Kronecker structure
(Hensman et al., 2013; Gal et al., 2014; Wilson & Nickisch,
2015; Hoang et al., 2016; Peng et al., 2017). A main
drawback of sparse approximations, however, is that the
representational capability is limited by the number of in-
ducing points (Moore & Russell, 2015). For example, for
a quick-varying function, the sparse approximations need
many inducing points to capture the local structures. That
is, this kind of scheme has not reduced the scaling of the
complexity (Bui & Turner, 2014).

The method exploited in this article belongs to the aggre-
gation models (Hinton, 2002; Tresp, 2000; Cao & Fleet,
2014; Deisenroth & Ng, 2015; Rullière et al., 2017), also
known as consensus statistical methods (Genest & Zidek,
1986; Ranjan & Gneiting, 2010). This kind of scheme produces
the final predictions by the aggregation of $M$ sub-
models (GP experts) respectively trained on the subsets
$\{D_i = \{X_i, y_i\}\}_{i=1}^M$ of $D = \{X, y\}$, thus distributing
the computations to “local” experts. Particularly, due to the
product of experts, the aggregation scheme derives a
factorized marginal likelihood for efficient training; and
then it combines the experts’ posterior distributions accord-
ing to a certain aggregation criterion. In comparison to
sparse approximations, the aggregation models (i) operate
directly on the full training data, (ii) require no additional
Generalized Robust Bayesian Committee Machine for Large-scale Gaussian Process Regression

inducing or variational parameters and (iii) distribute the computations on individual experts for straightforward parallelization (TavassoliPour et al., 2017), thus scaling them to arbitrarily large training data. In comparison to typical local GPs (Snelson & Ghahramani, 2007; Park et al., 2011), the aggregations smooth out the ugly discontinuity by the product of posterior distributions from GP experts. Note that the aggregation methods are different from the mixture-of-experts (Rasmussen & Ghahramani, 2002; Yuan & Neubauer, 2009), which suffers from intractable inference and is mainly developed for non-stationary regression.

However, it has been pointed out (Rullière et al., 2017) that there exists a particular type of training data such that typical aggregations, e.g., product-of-experts (PoE) (Hinton, 2002; Cao & Fleet, 2014) and Bayesian committee machine (BCM) (Tresp, 2000; Deisenroth & Ng, 2015), cannot offer consistent predictions, where “consistent” means the aggregated predictive distribution can converge to the true underlying predictive distribution when the training size $n$ approaches infinity.

The major contributions of this paper are three-fold. We first prove the inconsistency of typical aggregation models, e.g., the overconfident or conservative prediction variances illustrated in Fig. 3, using conventional disjoint or random data partition. Thereafter, we present a consistent yet efficient aggregation model for large-scale GP regression. Particularly, the proposed generalized robust Bayesian committee machine (GRBCM) selects a global subset to communicate with the remaining subsets, leading to the consistent aggregated predictive distribution derived under the Bayesian rule. Finally, theoretical and empirical analyses reveal that GRBCM outperforms existing aggregations due to the consistent yet efficient predictions. We release the demo codes in https://github.com/LiuHaiTao01/GRBCM.

2. Aggregation models revisited

2.1. Factorized training

A GP usually places a probability distribution over the latent function space as $f(x) \sim \mathcal{GP}(0, k(x, x'))$, which is defined by the zero mean and the covariance $k(x, x')$. The well-known squared exponential (SE) covariance function is

$$k(x, x') = \sigma_f^2 \exp\left(-\frac{1}{2} \sum_{i=1}^{d} \frac{(x_i - x'_i)^2}{l_i^2}\right),$$

where $\sigma_f^2$ is an output scale amplitude, and $l_i$ is an input length-scale along the $i$th dimension. Given the noisy observation $y(x) = f(x) + \epsilon$ where the i.i.d. noise follows $\epsilon \sim \mathcal{N}(0, \sigma_e^2)$ and the training data $D$, we have the marginal likelihood $p(y|X, \theta) = \mathcal{N}(0, k(X, X) + \sigma_e^2 I)$ where $\theta$ represents the hyperparameters to be inferred.

In order to train the GP on large-scale datasets, the aggregation models introduce a factorized training process. It first partitions the training set $D$ into $M$ subsets $D_i = \{X_i, y_i\}$, $1 \leq i \leq M$, and then trains GP on $D_i$ as an expert $\mathcal{M}_i$. In data partition, we can assign the data points randomly to the experts (random partition), or assign disjoint subsets obtained by clustering techniques to the experts (disjoint partition). Ignoring the correlation between the experts $\mathcal{M}_i$, $\mathcal{M}_j$ ($i \neq j$), the aggregated predictive distribution can converge to the true $\mathcal{M}_i$.

$$p(y|X, \theta) \approx \prod_{i=1}^{M} p_i(y_i|X_i, \theta),$$

where $p_i(y_i|X_i, \theta) \sim \mathcal{N}(0, K_i + \sigma_i^2 I_i)$ with $K_i = k(X_i, X_i) \in \mathbb{R}^{n_i \times n_i}$ and $n_i$ being the training size of $\mathcal{M}_i$. Note that for simplicity all the $M$ GP experts in (2) share the same hyperparameters as $\theta = \theta$ (Deisenroth & Ng, 2015). The factorization (2) degenerates the full covariance matrix $K = k(X, X)$ into a diagonal block matrix $\text{diag}(K_1, \ldots, K_M)$, leading to $K^{-1} \approx \text{diag}(K_1^{-1}, \ldots, K_M^{-1})$. Hence, compared to the full GP, the complexity of the factorized training process is reduced to $\mathcal{O}(mn^2)$ given $n_i = m_0 = n/M$, $1 \leq i \leq M$.

Conditioned on the related subset $D_i$, the predictive distribution $p_i(y_i|D_i, x_*) \sim \mathcal{N}(\mu_i(x_*), \sigma^2_i(x_*))$ of $\mathcal{M}_i$ has

$$\mu_i(x_*) = k_i^T [K_i + \sigma_i^2 I_i]^{-1} y_i, \quad (3a)$$
$$\sigma^2_i(x_*) = k(x_*, x_*) - k_i^T [K_i + \sigma_i^2 I_i]^{-1} k_i, \quad (3b)$$

where $k_i = k(X_i, x_*)$. Thereafter, the experts’ predictions $\{\mu_i, \sigma^2_i\}_{i=1}^{M}$ are combined by the following aggregation methods to perform the final predicting.

2.2. Prediction aggregation

The state-of-the-art aggregation methods include PoE (Hinton, 2002; Cao & Fleet, 2014), BCM (Tresp, 2000; Deisenroth & Ng, 2015), and nested pointwise aggregation of experts (NPAE) (Rullière et al., 2017).

For the PoE and BCM family, the aggregated prediction mean and precision are generally formulated as

$$\mu_A(x_*) = \sigma_A^2(x_*) \sum_{i=1}^{M} \beta_i \sigma_i^{-2}(x_*) \mu_i(x_*), \quad (4a)$$
$$\sigma_A^{-2}(x_*) = \sum_{i=1}^{M} \beta_i \sigma_i^{-2}(x_*) + (1 - \sum_{i=1}^{M} \beta_i) \sigma_e^{-2}, \quad (4b)$$

where the prior variance $\sigma_{**}^2 = k(x_*, x_*) + \sigma_e^2$, which is a correction term to $\sigma_A^{-2}$, is only available for the BCM family; and $\beta_i$ is the weight of the expert $\mathcal{M}_i$ at $x_*$. Instead of using $p_i(f_i|D_i, x_*)$ in (Deisenroth & Ng, 2015), we here consider the aggregations in a general scenario where each expert has all its belongings at hand.
The predictions of the PoE family, which omit the prior precision \( \sigma^2 \) in (4b), are derived from the product of \( M \) experts as

\[
p_A(y_*|D, x_*) = \prod_{i=1}^{M} p^\beta_i (y_*| D_i, x_*).
\]

The original PoE (Hinton, 2002) employs the constant weight \( \beta_i = 1 \), resulting in the aggregated prediction variances that vanish with increasing \( M \). On the contrary, the generalized PoE (GPoE) (Cao & Fleet, 2014) considers a varying weight \( \beta_i = 0.5(\log \sigma^2 + \log \sigma^2_* (x_*)), \) which represents the difference in the differential entropy between the prior \( p(y_*| x_*) \) and the posterior \( p(y_*| D_i, x_*) \), to weigh the contribution of \( M_i \) at \( x_* \). This varying \( \beta_i \) brings the flexibility of increasing or reducing the importance of experts based on the predictive uncertainty. However, the varying \( \beta_i \) may produce undesirable errors for GPoE. For instance, when \( x_* \) is far away from the training data such that \( \sigma^2 (x_*) \rightarrow \sigma^2_0, \) we have \( \beta_i \rightarrow 0 \) and \( \sigma^2_{GPoE} \rightarrow \infty \).

The BCM family, which is opposite to the PoE family, explicitly incorporates the GP prior \( p(y_*| x_*) \) when combining predictions. For two experts \( M_i \) and \( M_j \), BCM introduces a conditional independence assumption \( D_i \perp D_j|y_* \), leading to the aggregated predictive distribution as

\[
p_A(y_*| D, x_*) = \prod_{i=1}^{M} \frac{p^\beta_i (y_*| D_i, x_*)}{p^\beta_i (y_*| x_*)}.
\]

The original BCM (Tresp, 2000) employs \( \beta_i = 1 \) but its predictions suffer from weak experts when leaving the data. Hence, inspired by GPoE, the robust BCM (RBCM) (Deisenroth & Ng, 2015) uses a varying \( \beta_i \) to produce robust predictions by reducing the weights of weak experts. When \( x_* \) is far away from the training data \( X \), the correction term brought by the GP prior in (4b) helps the (R)BCM's prediction variance recover \( \sigma^2_* \). However, given \( M = 1 \), the predictions of RBCM as well as GPoE cannot recover the full GP predictions because usually \( \beta_i = 0.5(\log \sigma^2 + \log \sigma^2_* (x_*)) = 0.5(\log \sigma^2_* - \log \sigma^2_{full}(x_*)) \neq 1 \).

To achieve computation gains, the above aggregations introduce additional independence assumption for the experts' predictions, which however is often violated in practice and yields poor results. Hence, in the aggregation process, NPAE (Rullière et al., 2017) regards the prediction mean \( \mu_i (x_*) \) in (3a) as a random variable by assuming that \( y_* \) has not yet been observed, thus allowing for considering the covariances between the experts' predictions. Thereafter, for the random vector \( [\mu_1, \ldots, \mu_M, y_0]^T \), the covariances are derived as

\[
\text{cov}[\mu_i, y_*] = k_{i,0}^T k_{i,0} + k_{i,*}, \quad (7a)
\]

\[
\text{cov}[\mu_i, \mu_j] = \begin{cases} k_{i,0}^T k_{i,0} + k_{i,j} k_{j,0}^T k_{j,*}, & i \neq j, \\ k_{i,0}^T k_{i,0} + k_{i,j} k_{j,0}^T k_{j,*}, & i = j, \end{cases} \quad (7b)
\]

where \( K_{ij} = k(X_i, X_j) \in \mathbb{R}^{n_i \times n_j}, \) \( K_{i,0} = K_i + \sigma^2 I, \) \( K_{j,0} = K_j + \sigma^2 I, \) and \( K_{i,j} = K_{ij} + \sigma^2 I. \) With these covariances, a nested GP training process is performed to derive the aggregated prediction mean and variance as

\[
\begin{align}
\mu_{NPAE}(x_*) &= k_{A*}^T K_A^{-1} \mu, \\
\sigma^2_{NPAE}(x_*) &= k(x_*, x_*) - k_{A*}^T K_A^{-1} k_{A*} + \sigma^2, \quad (8a)
\end{align}
\]

where \( k_{A*} \in \mathbb{R}^{M \times 1} \) has the \( i \)th element as \( \text{cov} [\mu_i, y_*], \) \( K_A \in \mathbb{R}^{M \times M} \) has \( K_A^{ii} = \text{cov}[\mu_i, \mu_j], \) and \( \mu = [\mu_1(x_*), \ldots, \mu_M(x_*)]^T. \) The NPAE is capable of providing consistent predictions at the cost of implementing a much more time-consuming aggregation because of the inversion of \( K_A \) at each test point.

### 2.3. Discussions of existing aggregations

Though showcasing promising results (Deisenroth & Ng, 2015), given that \( n \rightarrow \infty \) and the experts are noise-free GPs, (GPoE and (R)BCM have been proved to be inconsistent, since there exists particular triangular array of data points that are dense in the input domain \( \Omega \) such that the prediction variances do not go to zero (Rullière et al., 2017).

Particularly, we further show below the inconsistency of (GPoE) and (R)BCM using two typical data partitions (random and disjoint partition) in the scenario where the observations are blurred with noise. Note that since GPoE using a varying \( \beta_i \) may produce undesirable errors, we adopt \( \beta_i = 1/M \) as suggested in (Deisenroth & Ng, 2015). Now the GPoE’s prediction mean is the same as that of PoE; but the prediction variance blows up as \( M \) times that of PoE.

**Definition 1.** When \( n \rightarrow \infty \), let \( X \in \mathbb{R}^{n \times d} \) be dense in \( \Omega \subset [0,1]^d \) such that for any \( x \in \Omega \) we have \( \lim_{n \rightarrow \infty} \min_{1 \leq i \leq n} \| x_i - x \| = 0. \) Besides, the underlying function to be approximated has true continuous response \( \mu(x) \) and true noise variance \( \sigma^2 \).

Firstly, for the disjoint partition that uses clustering techniques to partition the data \( D \) into disjoint local subsets \( \{ D_i \}_{i=1}^{M} \). The proposition below reveals that when \( n \rightarrow \infty \), PoE and (R)BCM produce overconfident prediction variance that shrinks to zero; on the contrary, GPoE provides conservative prediction variance.

**Proposition 1.** Let \( \{ D_i \}_{i=1}^{M} \) be a disjoint partition of the training data \( D \). Let the expert \( M_i \), trained on \( D_i \), be GP with zero mean and stationary covariance function \( k(\cdot, \cdot) > 0 \). We further assume that (i) \( \lim_{n \rightarrow \infty} M_n = \infty \) and (ii) \( \lim_{n \rightarrow \infty} n/M_n^2 > 0 \), where the second condition implies that the subset size \( M_n \) is comparable such that too weak experts are not preferred. Besides, from the second condition we have \( M_n \rightarrow \infty \), which implies that the experts become more informative with increasing \( n \). Then, PoE and (R)BCM
produce overconfident prediction variance at \( x_s \in \Omega \) as

\[
\lim_{n \to \infty} \sigma_{A,n}^2(x_s) = 0,
\]

whereas GPoE yields conservative prediction variance

\[
\sigma_n^2 < \lim_{n \to \infty} \sigma_{A,n}^2(x_s) < \sigma_n^2(x_s) < \sigma_{*2},
\]

where \( \sigma_n^2(x_s) \) is offered by the farthest expert \( M_n \) (1 \( \leq b_n \leq M_n \)) whose prediction variance is closest to \( \sigma_{*2}^2 \).

The detailed proof is given in Appendix A. Moreover, we have the following findings.

Remark 1. For the averaging \( \sigma_{GPoE}^2 = \frac{1}{M} \sum_{i=1}^{M} \sigma_{i}^2(x_s) \) and \( \mu_{(GPoE)} = \frac{1}{M} \sum_{i=1}^{M} \mu_i(x_s) \), using disjoint partition, more and more experts become relatively far away from \( x_s \) when \( n \to \infty \), i.e., the prediction variances at \( x_s \) approach \( \sigma_{*2}^2 \) and the prediction means approach the prior mean \( \mu_{*2} \). Hence, empirically, when \( n \to \infty \), the conservative \( \sigma_{GPoE}^2 \) illustrates an extremely time-consuming process of NPAE impose the demand of developing a consistent yet efficient aggregation model for large-scale GP regression.

As for data partition, intuitively, the random partition provides overlapping and coarse global information about the target function, which limits the ability to describe quick-varying characteristics. On the contrary, the disjoint partition provides separate and refined local information, which enables the model to capture the variability of target function. The superiority of disjoint partition has been empirically confirmed in (Rullière et al., 2017). Therefore, unless otherwise indicated, we employ disjoint partition for the aggregation models throughout the article.

As for time complexity, the five aggregation models have the same training process, and they only differ in how to combine the experts’ predictions. For (GPoE) and (R)BCM, their time complexity in prediction scales as \( O(nm^2) + O(n^\prime m_0) \) where \( n^\prime \) is the number of test points.\(^2\) For the complicated NPAE, it however needs to invert an \( M \times M \) matrix \( K_n \) at each test point, leading to a greatly increased time complexity in prediction as \( O(n^\prime n^2) \).\(^3\)

The inconsistency of (GPoE) and (R)BCM and the extremely time-consuming process of NPAE impose the demand of developing a consistent yet efficient aggregation model for large-scale GP regression.

3. Generalized robust Bayesian committee machine

3.1. GRBCM

Our proposed GRBCM divides \( M \) experts into two groups. The first group has a global communication expert \( M_c \) trained on the subset \( D_c = D_1 \), and the second group contains the remaining \( M - 1 \) global or local experts\(^4\) \( \{M_i\}_{i=2}^{M} \) trained on \( \{D_i\}_{i=2}^{M} \), respectively. The training process of GRBCM is identical to that of typical aggregations in section 2.1. The prediction process of GRBCM, however, is different. Particularly, GRBCM assigns the global communication expert with the following properties:

- (Random selection) The communication subset \( D_c \) is a random subset wherein the points are randomly selected.

\(^2\)\(O(nm^2)\) is induced by the update of \( M \) GP experts after optimizing hyperparameters.

\(^3\)The predicting complexity of NPAE can be reduced by employing various hierarchical computing structure (Rullière et al., 2017), which however cannot provide identical predictions.

\(^4\)Global” means the expert is trained on a random subset, whereas “local” means it is trained on a disjoint subset.
With (12), GRBCM’s predictive distribution is 

\[ p(x_t|D, x_*) \sim \mathcal{N}(\mu_{x_t}, \sigma^2_{x_t}) \]

Different from (R)BCM, GRBCM employs the informative \( \sigma^{-2}_{x_t} \) rather than the prior \( \sigma^{-2}_{x_t}^* \) to correct the prediction precision in (14b), leading to consistent predictions when \( n \to \infty \), which will be proved below. Also, the prediction mean of GRBCM in (14a) now is corrected by \( \mu_{x_t}(x_*) \).

**Proposition 3.** Let \( \{D_i\}_{i=1}^{M} \) be a partition of the training data \( D \) with (i) \( \lim_{n \to \infty} M_n = \infty \) and (ii) \( \lim_{n \to \infty} n/M_n > 0 \). Besides, among the \( M \) subsets, there is a global communication subset \( D_c \), the points in which are randomly selected from \( D \) without replacement. Let the global expert \( \mathcal{M}_c \) and the enhanced experts \( \{\mathcal{M}_{i+1}\}_{i=1}^{M} \) be GPs with zero mean and stationary covariance function \( k(.) \). Then, GRBCM yields consistent predictions as

\[
\lim_{n \to \infty} \mu_{GRBCM}(x_*) = \mu_{x_t}(x_*), \\
\lim_{n \to \infty} \sigma^2_{GRBCM}(x_*) = \sigma^2_{x_t}.
\]
and 3, both GPoE and GRBCM produce consistent predictions using random partition. It is known that the GP model $\mathcal{M}$ provides more confident predictions, i.e., lower uncertainty $U(\mathcal{M}) = \int \sigma^2(x)dx$, with more data points. Since GRBCM trains experts on more informative subsets $(\mathcal{D}_{i+1})_{i=1}^M$, we have the following finding.

**Remark 3.** When using random subsets, the GRBCM’s prediction uncertainty is always lower than that of GPoE, since the discrepancy $\delta_{U^{-1}} = U^{-1}_{\text{GRBCM}} - U^{-1}_{\text{GPoE}}$ satisfies

$$\delta_{U^{-1}} = \left[ U^{-1}(\mathcal{M}_{+2}) - \frac{1}{M} \sum_{i=1}^{M_i} U^{-1}(\mathcal{M}_i) \right]$$

$$+ \int \sum_{i=3}^{M_i} \beta_i \left( \sigma^2_{i+1}(x_*) - \sigma^2_{c}(x_*) \right) dx_* > 0$$

for a large enough $n$. It means compared to GPoE, GRBCM converges faster to the underlying function when $n \rightarrow \infty$.

Finally, similar to RBCM, GRBCM can be executed in multi-layer computing architectures with identical predictions (Deisenroth & Ng, 2015; Ionescu, 2015), which allow to run optimally and efficiently with the available computing infrastructure for distributed computing.

### 3.2. Complexity

Assuming that the experts $\{\mathcal{M}_i\}_{i=1}^{M}$ have the same training size $n_i = m_0 = n/M$ for $1 \leq i \leq M$. Compared to (G)PoE and (R)BCM, the proposed GRBCM has a higher time complexity in prediction due to the construction of new experts $\{\mathcal{M}_i\}_{i=2}^{M}$. In prediction, it first needs to calculate the inverse of $k(\mathbf{X}, \mathbf{X}_c)$ and $M-1$ augmented covariance matrices $\{k(\mathbf{X}_i, \mathbf{X}); \{\mathbf{X}_i, \mathbf{X}_c\}\}_{i=2}^{M}$, which scales as $O(8mn_0^3 - 7m_0^3)$, in order to obtain the predictions $\mu_c$, $\mu_{i+1}^{M_i} = 2$ and $\sigma^2_c$, $\sigma^2_{i+1}^{M_i} = 2$. Then, it combines the predictions of $\mathcal{M}_c$ and $\mathcal{M}_{i+1}^{M_i}$ at $n'$ test points. Therefore, the time complexity of the GRBCM prediction process is $O((nM_0^3 + 3nM_0))$, where $\alpha = (8M - 7)/M$ and $\beta = (4M - 3)/M$.

### 4. Numerical experiments

#### 4.1. Toy example

We employ a 1D toy example

$$f(x) = 5x^2 \sin(12x) + (x^3 - 0.5) \sin(3x - 0.5)$$

$$+ 4 \cos(2x) + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, 0.25)$, to illustrate the characteristics of existing aggregation models.

We generate $n = 10^4$, $5 \times 10^4$, $10^5$, $5 \times 10^5$ and $10^6$ training points, respectively, in $[0, 1]$, and select $n'$ = $0.1n$ test points randomly in $[-0.2, 1.2]$. We pre-normalize each column of $X$ and $y$ to zero mean and unit variance. Due to the global expert $\mathcal{M}_c$ in GRBCM, we slightly modify the disjoint partition: we first generate a random subset and then use the $k$-means technique to generate $M - 1$ disjoint subsets. Each expert is assigned with $m_0 = 500$ data points. We implement the aggregations by the GPML toolbox\(^2\) using the SE kernel in (1) and the conjugate gradients algorithm with the maximum number of evaluations as 500, and execute the code on a workstation with four 3.70 GHz cores and 16 GB RAM (multi-core computing in Matlab is employed). Finally, we use the Standardized Mean Square Error (SMSE) to evaluate the accuracy of prediction mean, and the Mean Standardized Log Loss (MSLL) to quantify the quality of predictive distribution (Rasmussen & Williams, 2006).

![Figure 2. Comparison of different aggregation models on the toy example in terms of (a) computing time, (b) SMSE and (c) MSLL.](http://www.gaussianprocess.org/gpml/code/matlab/doc/)

Fig. 2 depicts the comparative results of six aggregation models on the toy example. Note that NPAE using $n > 5 \times 10^4$ is unavailable due to the time-consuming prediction process. Fig. 2(a) shows that these models require the same training time, but they differ in the predicting time. Due to the communication expert, the GRBCM’s prediction time slightly offsets the curves of (G)PoE and (R)BCM. The NPAE however exhibits significantly larger predicting time with increasing $M$ and $n'$. Besides, Fig. 2(b) and (c) reveal that GRBCM and NPAE yield better predictions with increasing $n$, which confirm their consistency when $n \rightarrow \infty$.\(^6\) As for NPAE, though performing slightly better than GRBCM using $n = 5 \times 10^4$, it requires several orders of magnitude larger predicting time, rendering it unsuitable for cases with many test points and subsets.

Fig. 3 illustrates the six aggregation models using $n = 10^4$ and $n = 5 \times 10^5$, respectively, in comparison to the full GP (ground truth) using $n = 10^4$.\(^7\) It is observed that in terms of

\(^2\)http://www.gaussianprocess.org/gpml/code/matlab/doc/

\(^6\)Further discussions of GRBCM is shown in Appendix D.

\(^7\)The full GP is intractable using our computer for $n = 5 \times 10^5$. 

![Figure 3. Comparison of different aggregation models on the toy example in terms of (a) computing time, (b) SMSE and (c) MSLL.](http://www.gaussianprocess.org/gpml/code/matlab/doc/)
of prediction mean, as discussed in remark 1, PoE and GPoE provide poorer results in the entire domain with increasing \( n \). On the contrary, BCM and RBCM provide good predictions in the range \([0, 1]\). As discussed in remark 2, BCM however yields unreliable predictions when leaving the training data. RBCM alleviates the issue by using a varying \( \beta \). In terms of prediction variance, with increasing \( n \), PoE and (R)BCM tend to shrink to zero (overconfident), while GPoE tends to approach \( \sigma_{\text{post}}^2 \) (too conservative). Particularly, PoE always has the largest MSLL value in Fig. 2(b), since as discussed in remark 2, its prediction variance approaches zero faster.

### 4.2. Medium-scale datasets

We use two realistic datasets, \( \text{kin40k} \) (8D, \( 10^4 \) training points, \( 3 \times 10^4 \) test points) (Seeger et al., 2003) and \( \text{sarcos} \) (21D, 44484 training points, 4449 test points) (Rasmussen & Williams, 2006), to assess the performance of our approach.

The comparison includes all the aggregations except the expensive NPAE.\(^8\) Besides, we employ the fully independent training conditional (FITC) (Snelson & Ghahramani, 2006), the GP using stochastic variational inference (SVI)\(^9\) (Hensman et al., 2013), and the subset-of-data (SOD) (Chalupka et al., 2013) for comparison. We select the inducing size \( m \) for FITC and SVI, the batch size \( m_b \) for SOD, and the subset size \( m_{\text{sod}} \) for SOD, such that the computing time is similar to or a bit larger than that of GRBCM. Particularly, we choose \( m = 200, m_b = 0.1n \) and \( m_{\text{sod}} = 2500 \) for \( \text{kin40k} \), and \( m = 300, m_b = 0.1n \) and \( m_{\text{sod}} = 3000 \) for \( \text{sarcos} \). Differently, SVI employs the stochastic gradients algorithm with \( t_{\text{sg}} = 1200 \) iterations. Finally, we adopt the disjoint partition used before to divide the \( \text{kin40k} \) dataset into 16 subsets, and the \( \text{sarcos} \) dataset into 72 subsets for the aggregations. Each experiment is repeated ten times.

\(^8\)The comparison of NPAE and GRBCM are separately provided in Appendix E.

\(^9\)https://github.com/SheffieldML/GPy

Fig. 4 depicts the comparative results of different approximation models over 10 runs on the \( \text{kin40k} \) and \( \text{sarcos} \) datasets.

Fig. 4. Comparison of the approximation models on the \( \text{kin40k} \) and \( \text{sarcos} \) datasets.

Next, we explore the impact of the number \( M \) of experts on the performance of aggregations. To this end, we run them on the \( \text{kin40k} \) dataset with \( M \) respectively being 8, 16 and 64, and we run on the \( \text{sarcos} \) dataset with \( M \) respectively being 36, 72 and 288. The results in Fig. 5 turn out that all the aggregations perform worse with increasing \( M \), since the experts become weaker; but GRBCM still yields the best performance with different \( M \). Besides, with increasing \( M \), the poor prediction mean and the vanishing prediction variance of PoE result in the sharp increase of MSLL values.
This section explores the performance of aggregations and SVI on two large-scale datasets. We first assess them on the 90D song dataset, which is a subset of the million song dataset (Bertin-Mahieux et al., 2011). The song dataset is partitioned into 450000 training points and 65345 test points. We then assess the models on the 11D electric dataset that is partitioned into 1.8 million training points and 249280 test points. We follow the normalization and data pre-processing in (Wilson et al., 2016) to generate the two datasets.\footnote{The datasets and the pre-processing scripts are available in https://people.orie.cornell.edu/andrew/.

Table 1. Comparative results of the aggregation models and SVI on the song and electric datasets.

<table>
<thead>
<tr>
<th></th>
<th>song (450K)</th>
<th>electric (1.8M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMSE</td>
<td>MSSE</td>
<td>SMSE</td>
</tr>
<tr>
<td>PoE</td>
<td>0.8527</td>
<td>328.82</td>
</tr>
<tr>
<td>GPoE</td>
<td>0.8527</td>
<td>0.1159</td>
</tr>
<tr>
<td>BCM</td>
<td>2.6919</td>
<td>156.62</td>
</tr>
<tr>
<td>RBCM</td>
<td>1.3383</td>
<td>24.930</td>
</tr>
<tr>
<td>SVI</td>
<td>0.7909</td>
<td>-0.1885</td>
</tr>
<tr>
<td>GRBCM</td>
<td>0.7321</td>
<td>-0.1571</td>
</tr>
</tbody>
</table>

The superiority of local experts is the capability of capturing non-stationary and heteroscedastic features.}

Figure 5. Comparison of the aggregation models using different numbers of experts on the kin40k and sarcos datasets.

Finally, we investigate the impact of data partition (disjoint or random) on the performance of aggregations. The average results in Fig. 6 turn out that the disjoint partition is more beneficial for the aggregations. The results are expectable since the disjoint subsets provide separate and refined local information, whereas the random subsets provide overlapping and coarse global information. But we observe that GPoE performs well on the sarcos dataset using random partition, which confirms the conclusions in Proposition 2. Besides, as revealed in remark 3, even using random partition, GRBCM outperforms GPoE.

4.3. Large-scale datasets

This section explores the performance of aggregations and SVI on two large-scale datasets. We first assess them on the 90D song dataset, which is a subset of the million song dataset (Bertin-Mahieux et al., 2011). The song dataset is partitioned into 450000 training points and 65345 test points. We then assess the models on the 11D electric dataset that is partitioned into 1.8 million training points and 249280 test points. We follow the normalization and data pre-processing in (Wilson et al., 2016) to generate the two datasets.\footnote{The datasets and the pre-processing scripts are available in https://people.orie.cornell.edu/andrew/.

Table 1. Comparative results of the aggregation models and SVI on the song and electric datasets.

<table>
<thead>
<tr>
<th></th>
<th>song (450K)</th>
<th>electric (1.8M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMSE</td>
<td>MSSE</td>
<td>SMSE</td>
</tr>
<tr>
<td>PoE</td>
<td>0.8527</td>
<td>328.82</td>
</tr>
<tr>
<td>GPoE</td>
<td>0.8527</td>
<td>0.1159</td>
</tr>
<tr>
<td>BCM</td>
<td>2.6919</td>
<td>156.62</td>
</tr>
<tr>
<td>RBCM</td>
<td>1.3383</td>
<td>24.930</td>
</tr>
<tr>
<td>SVI</td>
<td>0.7909</td>
<td>-0.1885</td>
</tr>
<tr>
<td>GRBCM</td>
<td>0.7321</td>
<td>-0.1571</td>
</tr>
</tbody>
</table>

The superiority of local experts is the capability of capturing non-stationary and heteroscedastic features.}

Figure 6. Comparison of the aggregation models using disjoint and random partitions on the kin40k dataset (M = 16) and the sarcos dataset (M = 72).

5. Conclusions

To scale the standard GP to large-scale regression, we present the GRBCM aggregation model, which introduces a global communication expert to yield consistent yet efficient predictions when $n \to \infty$. Through theoretical and empirical analyses, we demonstrated the superiority of GRBCM over existing aggregations on datasets with up to 1.8M training points. The superiority of local experts is the capability of capturing local patterns. Hence, further works will consider the experts with individual hyperparameters in order to capture non-stationary and heteroscedastic features.
Acknowledgements

This work was conducted within the Rolls-Royce@NTU Corporate Lab with support from the National Research Foundation (NRF) Singapore under the Corp Lab@University Scheme. It is also partially supported by the Data Science and Artificial Intelligence Research Center (DSAIR) and the School of Computer Science and Engineering at Nanyang Technological University.

References


A. Proof of Proposition 1

With disjoint partition, we consider two extreme local GP experts. For the first extreme expert $M_{a_n}$, the test point $x_*$ falls into the local region defined by $X_n$, i.e., $x_*$ is adherent to $X_{a_n}$ when $n \to \infty$. Hence, we have (Vazquez & Bect, 2010)

$$\lim_{n \to \infty} \sigma_{a_n}^2(x_*) = \lim_{n \to \infty} \sigma_{e_n}^2 = \sigma_{y}^2.$$  (18)

For the other extreme expert $M_{b_n}$, it lies farthest away from $x_*$ such that the related prediction variance $\sigma_{b_n}^2(x_*)$ is closest to $\sigma_{\ast\ast}^2$. It is known that for any $M_i$ ($i \neq a_n$) where $x_*$ is away from the training data $X_i$, given the relative distance $r_i = \min ||x_* - x||_{\forall x \in X_i}$, we have $\lim_{r_i \to \infty} \sigma_i^2(x_*) = \sigma_{\ast\ast}^2$. Since, however, we here focus on the GP predictions in the bounded region $\Omega \in [0,1]^d$ and employ the covariance function $k(.) > 0$, then the positive sequence $c_n = \{\sigma_{b_n}^2(x_*) - \sigma_{\ast\ast}^2\}$ is small but satisfies $\lim_{n \to \infty} c_n > 0$ and

$$\sigma_i^{-2}(x_*) - \sigma_{\ast\ast}^{-2} \geq c_n, \quad 1 \leq i \neq a_n \leq M_n.$$  

The equality holds only when $i = b_n$.

Thereafter, with the sequence $\epsilon_n = \min\{c_n, \frac{1}{M_n}\} \to \infty$ where $\alpha > 0$ we have

$$\sigma_i^{-2}(x_*) - \sigma_{\ast\ast}^{-2} \geq \epsilon_n \geq c_n, \quad 1 \leq i \neq a_n \leq M_n.$$  

It is found that $\epsilon_n = c_n$ is possible to hold only when $M_n$ is small. With the increase of $n$, $\epsilon_n$ quickly becomes much smaller than $c_n$ since $\lim_{n \to \infty} \epsilon_n \to \infty$.

The typical aggregated prediction variance writes

$$\sigma_{A,n}^{-2}(x_*) = \sum_{i=1}^{M_n} \beta_i (\sigma_i^{-2}(x_*) - \sigma_{\ast\ast}^{-2}) + \sigma_{\ast\ast}^{-2},$$  (18)

where for (G)PoE we remove the prior precision $\sigma_{\ast\ast}^{-2}$. We prove below the inconsistency of (G)PoE and (R)BCM using disjoint partition.

For PoE, (18) is $\sum_{i=1}^{M_n} \sigma_i^{-2}(x_*) > M_n \sigma_{\ast\ast}^{-2} \to n \to \infty$, leading to the inconsistent variance $\lim_{n \to \infty} \sigma_{A,n}^{-2} = 0$. For (R)BCM, the first term of $\sigma_{A,n}^{-2}(x_*)$ in (18) satisfies, given
that $n$ is large enough,
\[ \sum_{i=1}^{M_n} \beta_i (\sigma_i^2(x_i) - \sigma_{\ast}^2) > \epsilon_n \sum_{i=1}^{M_n} \beta_i = \frac{1}{M_n} \sum_{i=1}^{M_n} \beta_i. \]
Taking $\beta_i = 1$ for BCM and $\alpha = 0.5$, we have
\[ \frac{1}{M_n} \sum_{i=1}^{M_n} \beta_i = \sqrt{M_n} \rightarrow \infty \text{ as } n \rightarrow \infty, \]
leading to the inconsistent variance $\lim_{n \rightarrow \infty} \sigma_{\ast,i}^2 = 0$. For RBCM, since
\[ \beta_i = 0.5 (\log \sigma_{\ast,i}^2 - \log \sigma_i^2(x_i)) \geq 0.5 \log (1 + c_n \sigma_{\ast,i}^2), \]
where the equality holds only when $i = b_n$, we have
\[ \frac{1}{M_n} \sum_{i=1}^{M_n} \beta_i > 0.5 \log (1 + c_n \sigma_{\ast,i}^2) \sqrt{M_n} \rightarrow \infty \text{ as } n \rightarrow \infty, \]
leading to the inconsistent variance $\lim_{n \rightarrow \infty} \sigma_{\ast,i}^2 = 0$.

Finally, for GPoE, we know that when $n \rightarrow \infty$, $\sigma_{\ast,i}^2(x_i)$ converges to $\sigma_{\ast}^2$; but the other prediction precisions satisfy
\[ \epsilon_n + \sigma_{\ast,i}^2 \leq \sigma_{\ast,i}^2(x_i) < \sigma_{\ast,i}^2 \rightarrow \infty \text{ for } 1 \leq i \neq a_n \leq M_n, \]
since $x_i$ is away from their training points. Hence, we have
\[
\lim_{n \rightarrow \infty} (\sigma_{\ast,i}^2 - \sigma_{\ast,n}^2(x_i)) = \frac{1}{M_n} \left( \sigma_{\ast,i}^2 - \sigma_{\ast,n}^2(x_i) \right)
+ \lim_{n \rightarrow \infty} \frac{1}{M_n} \sum_{i \neq a_n} \left( \sigma_{\ast,i}^2 - \sigma_{\ast,n}^2(x_i) \right)
> \lim_{n \rightarrow \infty} \frac{1}{M_n} \left( \sigma_{\ast,i}^2 - \sigma_{\ast,n}^2(x_i) \right)
+ \lim_{n \rightarrow \infty} \frac{1}{M_n} \sum_{i \neq a_n} \left( \sigma_{\ast,i}^2 - \sigma_{\ast,n}^2(x_i) \right) = 0,
\]
which means that $\sigma_{\ast,n}^2(x_i)$ is inconsistent since
\[ \lim_{n \rightarrow \infty} \sigma_{\ast,n}^2(x_i) > \sigma_{\ast}^2. \]
Meanwhile, we easily find that $\lim_{n \rightarrow \infty} \sigma_{\ast,n}^2(x_i) = \epsilon_n + \sigma_{\ast,i}^2$, leading to
\[ \lim_{n \rightarrow \infty} \sigma_{\ast,n}^2(x_i) < \sigma_{\ast,i}^2(x_i) < \sigma_{\ast,i}^2. \]

**B. Proof of Proposition 2**

With smoothness assumption and particularly distributed noise (normal or Laplacian distribution), it has been proved that the GP predictions would converge to the true predictions when $n \rightarrow \infty$ (Choi & Schervish, 2004). Hence, given that the points in $X_i$ are randomly selected without replacement from $X$ and $n_i = n/M_n \rightarrow n \rightarrow \infty$, we have
\[ \lim_{n \rightarrow \infty} \mu_i(x_i) = \mu_n(x_i), \quad \lim_{n \rightarrow \infty} \sigma_i^2(x_i) = \sigma_n^2, \quad 1 \leq i \leq M_n. \]

For the aggregated prediction variance, we have
\[ \lim_{n \rightarrow \infty} \sigma_{\ast,n}^2(x_i) = \lim_{n \rightarrow \infty} \left( \sum_{i=1}^{M_n} \beta_i (\sigma_i^2(x_i) - \sigma_{\ast}^2) + \sigma_{\ast}^2 \right), \]
where for (G)PoE we remove $\sigma_{\ast,n}^2$. For PoE, given $\beta_i = 1$ and $\lim_{n \rightarrow \infty} \sigma_i^2(x_i) = \sigma_n^2$, we have the inconsistent variance $\lim_{n \rightarrow \infty} \sigma_{\ast,n}^2(x_i) = \lim_{n \rightarrow \infty} M_n \sigma_{\ast}^2 = \infty$. For GPoE, given $\beta_i = 1/M_n$ we have the consistent variance $\lim_{n \rightarrow \infty} \sigma_{\ast,n}^2(x_i) = M_n \sigma_{\ast}^2 = \sigma_n^2$. Finally, for RBCM, given $\beta_i = 1$ we have the consistent variance $\lim_{n \rightarrow \infty} \sigma_{\ast,n}^2(x_i) = \lim_{n \rightarrow \infty} M_n \sigma_{\ast}^2 = \sigma_n^2$.

Then, for the aggregated prediction mean we have
\[ \lim_{n \rightarrow \infty} \mu_{\ast,n}(x_i) = \lim_{n \rightarrow \infty} \sigma_{\ast,n}^2(x_i) \sum_{i=1}^{M_n} \beta_i (\sigma_i^2(x_i) \mu_i(x_i)). \]

For PoE, given $\beta_i = 1$ and $\lim_{n \rightarrow \infty} \sigma_i^2(x_i)/\sigma_n^2 = 1/M_n$, we have the consistent prediction mean $\lim_{n \rightarrow \infty} \mu_{\ast,n}(x_i) = \mu_n(x_i)$. For GPoE, given $\beta_i = 1/M_n$ and $\lim_{n \rightarrow \infty} \sigma_i^2(x_i)/\sigma_n^2 = 1$, we have the consistent prediction mean $\lim_{n \rightarrow \infty} \mu_{\ast,n}(x_i) = \mu_n(x_i)$. For (R)BCM, given $\beta_i = \beta_n = 1$ or $\lim_{n \rightarrow \infty} \beta_i = \beta_n = 0.5 \log (\sigma_n^2/\sigma_i^2)$, we have the inconsistent prediction mean $\lim_{n \rightarrow \infty} \mu_{\ast,n}(x_i) = \lim_{n \rightarrow \infty} \beta_n \sigma_n^2 \mu_n(x_i)/(\beta_n \sigma_n^2 - \sigma_i^2) + \sigma_i^2/M_n = a \mu_n(x_i)$ where $a = \sigma_n^2/\sigma_i^2$ and the equality holds when $\sigma_i^2 = 0$.

**C. Proof of Proposition 3**

Given that the points in the communication subset $D_c$ are randomly selected without replacement from $D$ and $n_c = n/M_n \rightarrow n \rightarrow \infty$, we have $\lim_{n \rightarrow \infty} \mu_{\ast}(x_i) = \mu_n(x_i)$ and $\lim_{n \rightarrow \infty} \sigma_{\ast,n}^2(x_i) = \sigma_n^2$ for $M_n$. Likewise, for the expert $M_{+1}$ trained on the augmented dataset $D_{+1} = \{D_1, D_c\}$ with size $n_{+1} = 2n/M_n$ we have $\lim_{n \rightarrow \infty} \mu_{+1}(x_i) = \mu_n(x_i)$ and $\lim_{n \rightarrow \infty} \sigma_{+1,n}^2(x_i) = \sigma_n^2$ for $2 \leq i \leq M$.

We first derive the upper bound of $\sigma_n^2(x_i)$. For the stationary covariance function $k(.) > 0$, when $n_c$ is large enough we have (Vazquez & Bect, 2010)
\[ \sigma_n^2(x_i) \leq k(x_i, x_a) - k(x_i, x_a') + \frac{\sigma_{\ast,n}^2(x_i)}{k(x_i, x_a')}, \]
where $x' \in X_c$ is the nearest data point to $x_i$. It is known that the relative distance $r_a = ||x_i - x_a||$ is proportional to the inverse of the training size $n_c$, i.e., $r_a \propto 1/n_c = M_n/n \rightarrow n \rightarrow \infty$. Conventional stationary covariance functions only rely on the relative distance (once the covariance parameters have been determined) and decrease with $r_c$. Consequently, the prediction variance $\sigma_n^2(x_i)$ increases with $r_c$. Taking the SE covariance function in (1) for ex-
ample,\(^\text{11}\) when \(r_c \to 0\) we have, given \(l_0 = \min_{1 \leq i \leq d} \{l_i\}\),
\[
\begin{align*}
\sigma^2(x_*) &\leq \sigma^2_1 - \sigma^2_2 \exp(-r^2_c/l_0^2) + \sigma^2_{c,n} \\
&< \frac{\sigma^2_2}{l_0} r^2_c + \sigma^2_{c,n} = \sigma^2_{c,n}.
\end{align*}
\]
(19)

We clearly see from this inequality that when \(r_c \to 0\), \(\sigma^2_{c}(x_*)\) goes to \(\sigma^2_{c}\) since \(\lim_{n \to \infty} \sigma^2_{c,n} = \sigma^2_{c}\).

Then, we rewrite the precision of GRBCM in (14b) as, given \(\beta_2 = 1\),
\[
\sigma^{-2}_{\text{GRBCM}}(x_*) = \sigma^{-2}_{+2}(x_*) + \sum_{i=3}^{M} \beta_i \left( \sigma^{-2}_{i+1}(x_*) - \sigma^{-2}_{c}(x_*) \right).
\]
(20)

Compared to \(M_c, M_{++}\) is trained on a more dense dataset \(D_{++}\), leading to \(\sigma^2_{i+1}(x_*) \leq \sigma^2_{c}(x_*)\) for a large enough \(n\).\(^\text{12}\)

Given (19) and \(\sigma^2_{i+1}(x_*) > \sigma^2_{c,n}\), the weight \(\beta_i\) satisfies, for \(3 \leq i \leq M_n\),
\[
0 \leq \beta_i = \frac{1}{2} \log \left( \frac{\sigma^2_{i+1}(x_*)}{\sigma^2_{+1}(x_*)} \right) < \frac{1}{2} \log \left( \frac{\sigma^2_{+1}(x_*)}{\sigma^2_{c,n}} \right) \leq \frac{a}{2r^2_c} \sigma_{c,n}^2.
\]
(21)

Besides, the precision discrepancy satisfies, for \(3 \leq i \leq M_n\),
\[
0 \leq \sigma^{-2}_{i+1}(x_*) - \sigma^{-2}_{c}(x_*) = \sigma^{-2}_{c}(x_*) \left( \frac{\sigma^2_{c}(x_*)}{\sigma^2_{+1}(x_*)} - 1 \right) \leq \frac{1}{\sigma^2_{c,n}} r^2_c.
\]
(22)

Hence, the second term in the right-hand side of (20) satisfies
\[
\sum_{i=3}^{M} \beta_i \left( \sigma^{-2}_{i+1}(x_*) - \sigma^{-2}_{c}(x_*) \right) < \sum_{i=3}^{M} \frac{a^2}{2r^2_c} n^4 \propto \frac{M_n^5}{n^4}.
\]
Since \(\lim_{n \to \infty} n/M_n^2 > 0\), we have \(\lim_{n \to \infty} n^4/M_n^5 = \infty\), and furthermore,
\[
\lim_{n \to \infty} \sum_{i=3}^{M} \beta_i \left( \sigma^{-2}_{i+1}(x_*) - \sigma^{-2}_{c}(x_*) \right) = 0.
\]
(23)

Substituting (23) and \(\lim_{n \to \infty} \sigma^{-2}_{i+1}(x_*) = \sigma_{c,n}^{-2}\) into (20), we have a consistent prediction precision as
\[
\lim_{n \to \infty} \sigma^{-2}_{\text{GRBCM}}(x_*) = \sigma_{c,n}^{-2}.
\]

11 We take the SE kernel for example since conventional kernels, e.g., the rational quadratic kernel and the Matern class of kernels, can reduce to the SE kernel under some conditions.

12 The equality is possible to hold when we employ disjoint partition for \(\{D_i\}_{i=2}^{M_n}\) and \(x_*\) is away from \(X_c\).

Similarly, we rewrite the GRBCM’s prediction mean in (14a) as
\[
\mu_{\text{GRBCM}}(x_*) = \sigma^2_{\text{GRBCM}}(x_*) \left( \mu_\Delta + \sigma^{-2}_{+2}(x_*) \mu_{+2}(x_*) \right),
\]
(24)

where
\[
\mu_\Delta = \sum_{i=3}^{M} \beta_i \left( \sigma^{-2}_{i+1}(x_*) \mu_{i+1}(x_*) - \sigma^{-2}_{c}(x_*) \mu_c(x_*) \right).
\]

Let \(\delta_{\text{max}} = \max_{3 \leq i \leq M_n} \left| \frac{\sigma^2_{i+1}(x_*)}{\sigma^2_{c,n}} \mu_{i+1}(x_*) - \mu_c(x_*) \right| \to n \to \infty 0\), we have
\[
|\mu_\Delta| \leq \sum_{i=3}^{M} \beta_i |\sigma^{-2}_{i+1}(x_*) \mu_{i+1}(x_*) - \mu_c(x_*)| \leq \sum_{i=3}^{M} \frac{a^2 r^2_c}{2\sigma^2_{c,n}} \delta_{\text{max}} \to n \to \infty 0.
\]
(25)

Substituting (25) into (24), we have the consistent prediction mean as
\[
\lim_{n \to \infty} \mu_{\text{GRBCM}}(x_*) = \mu_{\eta}(x_*).
\]

D. Discussions of GRBCM on the toy example

It is observed that the proposed GRBCM showcases superiority over existing aggregations on the toy example, which is brought by the particularly designed aggregation structure: the global communication expert \(M_c\) to capture the long-term features of the target function, and the remaining experts \(\{M_{++}\}_{i=2}^{M}\) to refine local predictions.

Figure 7. Comparative results of GRBCM and \(M_c\) on the toy example.

To verify the capability of GRBCM, we compare it with the pure global expert \(M_c\), which relies on a random subset \(X_c\). Fig. 7 shows the comparative results of GRBCM and \(M_c\) on the toy example. It is found that with increasing \(n\), (i) GRBCM always outperforms \(M_c\) because of the benefits brought by local experts; and (ii) the predictions of \(M_c\) generally become poorer since it becomes intractable to choose a good subset from the increasing dataset.

E. Experimental results of NPAE

Table 2 compares the results of GRBCM and NPAE over 10 runs on the \(kin40k\) dataset \((M = 16)\) and the \(sarcos\)
Table 2. Comparative results (mean and standard deviation) of GRBCM and NPAE over 10 runs on the *kin40k* dataset ($M = 16$) and the *sarcos* dataset ($M = 72$) using disjoint partition. The computing time $t$ for each model involves the training and predicting time.

<table>
<thead>
<tr>
<th></th>
<th>GRBCM</th>
<th>NPAE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>kin40k</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMSE</td>
<td>0.0223 ± 0.0005</td>
<td>0.0246 ± 0.0007</td>
</tr>
<tr>
<td>MSLL</td>
<td>-1.9927 ± 0.0177</td>
<td>-1.9565 ± 0.0170</td>
</tr>
<tr>
<td>$t$ [s]</td>
<td>78.1 ± 4.4</td>
<td>2852.4 ± 16.7</td>
</tr>
<tr>
<td><strong>sarcos</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMSE</td>
<td>0.0074 ± 0.0002</td>
<td>0.0054 ± 0.0001</td>
</tr>
<tr>
<td>MSLL</td>
<td>-2.3681 ± 0.0242</td>
<td>-2.5900 ± 0.0068</td>
</tr>
<tr>
<td>$t$ [s]</td>
<td>445.6 ± 49.4</td>
<td>26444.0 ± 1213.0</td>
</tr>
</tbody>
</table>

dataset ($M = 72$) using disjoint partition. It is observed that GRBCM performs slightly better than NPAE on the *kin40k* dataset, and produces competitive results on the *sarcos* dataset. But in terms of the computing efficiency, since NPAE needs to build and invert an $M \times M$ covariance matrix at each test point, it requires much more running time, especially for the *sarcos* dataset with $M = 72$. 