ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Fast Greedy Insertion and Deletion in Sparse Gaussian Process Regression

Jens Schreiter[1,2], Duy Nguyen-Tuong[1], Heiner Markert[1], Michael Hanselmann[1], and Marc Toussaint[2]

1- Robert Bosch GmbH - 70442 Stuttgart - Germany

2- University of Stuttgart - MLR Laboratory - 70569 Stuttgart - Germany

Abstract. In this paper, we introduce a new and straightforward criterion for successive insertion and deletion of training points in sparse Gaussian process regression. Our novel approach is based on an approximation of the selection technique proposed by Smola and Bartlett [1]. It is shown that the resulting selection strategies are as fast as the purely randomized schemes for insertion and deletion of training points. Experiments on real-world robot data demonstrate that our obtained regression models are competitive with the computationally intensive state-of-the-art methods in terms of generalization accuracy.

1 Motivation

Today, Gaussian processes are widely used non-parametric Bayesian modeling techniques [2]. However, the applicability of full Gaussian process regression (GPR) to large scale problems with a high number of training points n is limited due to the unfavourable scaling in training time and memory requirements. The dominating factors are usually $\mathcal{O}(n^3)$ costs for inversion of a dense covariance matrix $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ between all available training points and the $\mathcal{O}(n^2)$ space required to store it in memory. Furthermore, the full GPR model needs $\mathcal{O}(dn)$ costs for predicting a test instance, where d is the data dimension.

To overcome these limitations in computational costs and storage requirements, various sparse likelihood approximations have emerged recently, whose relations have been formalized in the unifying framework [3]. The fully independent training conditional (FITC) approximation [4] uses a flexible subset of virtual training points to generate a sparse GPR model and optimizes the virtual training points along with all other hyperparameters. In contrast, the deterministic training conditional (DTC) approximation selects a representative subset of real training points, the so-called active points, that induce the sparse approximation. Therefore, many greedy selection criteria were proposed, e.g. in [1, 5, 6, 7, 8, 9]. Many of these methods have significantly higher computational costs than randomized selection, but in exchange yield significantly better results, since random selection typically leads to over- or underfitting. In addition to a selection heuristic, Csató and Opper [9] introduced a highly similar heuristic for deletion of training points from the active set. They show that removing active points can considerably reduce the prediction times for test points with only slightly decreasing generalization accuracy. All of the insertion and deletion methods mentioned above either lack computational speed, have high memory

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

requirements, or lack of modeling accuracy. Moreover, if the regression model generation is based on a purely randomized selection or on a method with a small randomly selected subset of remaining training points for criteria evaluation, e.g. as done by [6, 8], the performance in hard regression tasks deteriorates.

Our newly developed method is closely related to the inclusion heuristic by Smola and Bartlett [1], but some reasonable assumptions reduce the computational costs to the level of randomized selection without a huge loss in model accuracy. Compared to the deletion criterion from Csató and Opper [9], our approach offers nearly the same prediction performance under considerably lower computing time.

2 Sparse Gaussian Process Regression

Let $\mathcal{D} = (\boldsymbol{y}, \boldsymbol{X})$ be the training data set, where $\boldsymbol{y} \in \mathbb{R}^n$ is a vector of noisy realizations of the underlying scalar function $f(\boldsymbol{x}_i) = f_i$, obeying the relationship $y_i = f_i + \varepsilon_i$ with Gaussian noise $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Furthermore, the n training inputs $\boldsymbol{x}_i \in \mathbb{R}^d$ are row-wise summarized in $\boldsymbol{X} \in \mathbb{R}^{n \times d}$. Our goal is the construction of a sparse GPR model which estimates the relationship above. Let I be the index set of size m of all active points $\boldsymbol{x}_i$ with $i \in I$, i.e. training points that represent the sparse approximation. As shown in Seeger et al. [5], through a centered Gaussian prior distribution with covariance matrix $\boldsymbol{K}_I \in \mathbb{R}^{m \times m}$ and an information optimal likelihood approximation with respect to the Kullback-Leibler divergence we get the approximated Gaussian posterior distribution

$$Q_I(\boldsymbol{f} \mid \boldsymbol{y}, \boldsymbol{X}) = \mathcal{N}\big(\boldsymbol{f} \;\big|\; \boldsymbol{V}^T \boldsymbol{L}_M^{-T} \boldsymbol{\beta}_I \,,\; \boldsymbol{K} - \boldsymbol{V}^T \boldsymbol{V} + \sigma^2 \boldsymbol{V}^T \boldsymbol{M}^{-1} \boldsymbol{V}\big) \tag{1}$$

for all training points with the estimated mean vector $\boldsymbol{\mu}_I = \boldsymbol{V}^T \boldsymbol{L}_M^{-T} \boldsymbol{\beta}_I \in \mathbb{R}^n$. Here, $\boldsymbol{L} \in \mathbb{R}^{m \times m}$ is the lower Cholesky factor of $\boldsymbol{K}_I$, $\boldsymbol{V} = \boldsymbol{L}^{-1} \boldsymbol{K}_{I,\cdot} \in \mathbb{R}^{m \times n}$, $\boldsymbol{M} = \sigma^2 \boldsymbol{I} + \boldsymbol{V}\boldsymbol{V}^T \in \mathbb{R}^{m \times m}$ with the Cholesky decomposition $\boldsymbol{M} = \boldsymbol{L}_M \boldsymbol{L}_M^T$, $\boldsymbol{\beta}_I = \boldsymbol{L}_M^{-1} \boldsymbol{V} \boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{\alpha}_I = \boldsymbol{L}^{-T} \boldsymbol{L}_M^{-T} \boldsymbol{\beta}_I \in \mathbb{R}^m$ for fixed I of size m. Due to the matrix-matrix multiplications, the training complexity of this sparse GPR model is $\mathcal{O}(nm^2)$. Predicting the mean for one test point is feasible in $\mathcal{O}(dm)$.

Most of the GP approximation techniques differ in the way, how the active set $\boldsymbol{X}_I$ is selected [3]. Usually, the remaining point that has the maximum gain with respect to an insertion criterion Δ_i is selected. One of the best selection methods is proposed by Smola and Bartlett [1]. They select the remaining point that maximizes the posterior likelihood for the admission of the prediction vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ of the full GP under the given data set $\mathcal{D}$. This is based on the transformation of $\boldsymbol{\alpha} = \boldsymbol{K}^{-1} \boldsymbol{f}$ and leads to the equivalent formulation

$$\tau_I = \min_{\boldsymbol{\alpha}_I} \left(\frac{1}{2} \boldsymbol{\alpha}_I^T \boldsymbol{L} \boldsymbol{M} \boldsymbol{L}^T \boldsymbol{\alpha}_I - \boldsymbol{\alpha}_I^T \boldsymbol{L} \boldsymbol{V} \boldsymbol{y} \right) = -\frac{1}{2} \boldsymbol{\beta}_I^T \boldsymbol{\beta}_I \tag{2}$$

in the sparse sense as pointed out in [5]. In the following, let $I' = I \cup \{i\}$. The decrease in the sparse posterior likelihood derived from (2) defines the selection

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

criterion by Smola and Bartlett (SB), i.e.

$$_{SB}\Delta_i = \tau_I - \tau_{I'} = \frac{1}{2}\beta_{I',i}^2 \ , \tag{3}$$

for a remaining point and with the new component $\beta_{I',i}$ of the updated vector $\boldsymbol{\beta}_{I'}$. Due to its high computational costs of $\mathcal{O}(nm)$ per remaining point for the criterion calculation, they only evaluate it for a randomly chosen subset of cardinality κ. The authors of [1] recommend $\kappa = 59$, which they justify with a probabilistic argument. Nevertheless, they end up with high computational costs of $\mathcal{O}\left(\kappa nm^2\right)$ for the whole DTC approximation. The conjugation of this selection heuristic defines also a corresponding deletion criterion $_{SB}\nabla_i$ which leads to $\mathcal{O}\left(m^2\right)$ costs per active point. Always the active point according to the posterior model (1) with minimal loss in terms of a deletion criterion is removed.

3 Maximum Error Criterion to Speed up Sparse GPR

In this section, we first discuss the successive inclusion of training points into the active subset. To include a remaining point $\boldsymbol{x}_i$ in the active subset, we have to update the Cholesky factors $\boldsymbol{L}$, $\boldsymbol{L}_M$, the matrix $\boldsymbol{V}$, respectively $\boldsymbol{K}_{I,\cdot}$, the vector $\boldsymbol{\beta}_I$, and the mean $\boldsymbol{\mu}_I$ of the posterior distribution (1), as shown in [5]. The costs for the sequential insertion in the m-th iteration are $\mathcal{O}(nm)$. Our approach maximizes also the evidence of the current posterior model, which is similar to the greedy scheme (3). This strategy leads to successively maximization of the Euclidean norm of the vector $\boldsymbol{\beta}_{I'}$, which is equivalent to iteratively minimizing $\|\boldsymbol{y} - \boldsymbol{\mu}_{I'}\|$ for the normalized vector $\boldsymbol{y}$ and thus approximately normalized $\boldsymbol{\mu}_{I'}$, since we have $\|\boldsymbol{\beta}_{I'}\|^2 = \boldsymbol{\beta}_{I'}^T\boldsymbol{\beta}_{I'} = \boldsymbol{y}^T\boldsymbol{\mu}_{I'}$ after an inclusion. Due to the equivalence of norms in finite dimensional spaces it holds true that $\|\boldsymbol{y} - \boldsymbol{\mu}_{I'}\| \leq n \max_{\forall j} \left|y_j - \mu_{I',j}\right|$. In the limit, i.e. with increasing m, we have $\boldsymbol{\mu}_{I'} \approx \boldsymbol{\mu}_I$. Thus, we define

$$_{\text{ME}}\Delta_i = \left|y_i - \mu_{I,i}\right| \tag{4}$$

as our new selection criterion and select the remaining point that has the maximal error (ME) under the current posterior model (1). This computational efficient approach has $\mathcal{O}(1)$ costs for criterion calculation per remaining point. The intuitive convergence assumption obviates the update of the posterior model for each remaining point as needed for the selection criterion (3).

In the following, we present our maximum error deletion criterion for the removal of active points. Typically, the maximum number of active points m is predefined and directly influences computing time (quadratically) and memory requirements (linearly). The deletion of appropriate active points improves the predictive performance without significantly deteriorating the existing model quality, e.g. see [9]. It also provides a way to reduce redundancy in the greedily selected active subset. Similar to the presented insertion strategy, we opt for a greedy criterion to successively delete active points. Note that deleting an active point does not necessarily lead to a state that was previously encountered in iteratively inserting training points. The reason is that the underlying assumptions

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

for greedy insertion and deletion differ considerably. While the inclusion strategy uses Cholesky updates, QR-downdates based on the factorization $\boldsymbol{QR} = \boldsymbol{LML}^T$ are used for deleting active points since they offer higher numerical stability. Inspired by the criterion of [9], we define our new deletion criterion as follows. Beginning with an already selected subset determined by I we remove the active point that has the minimal value with respect to the deletion criterion

$$_{\text{ME}}\nabla_i = \left|{}_{\text{ME}}\Delta_i \, \alpha_{I,i}\right| \ , \tag{5}$$

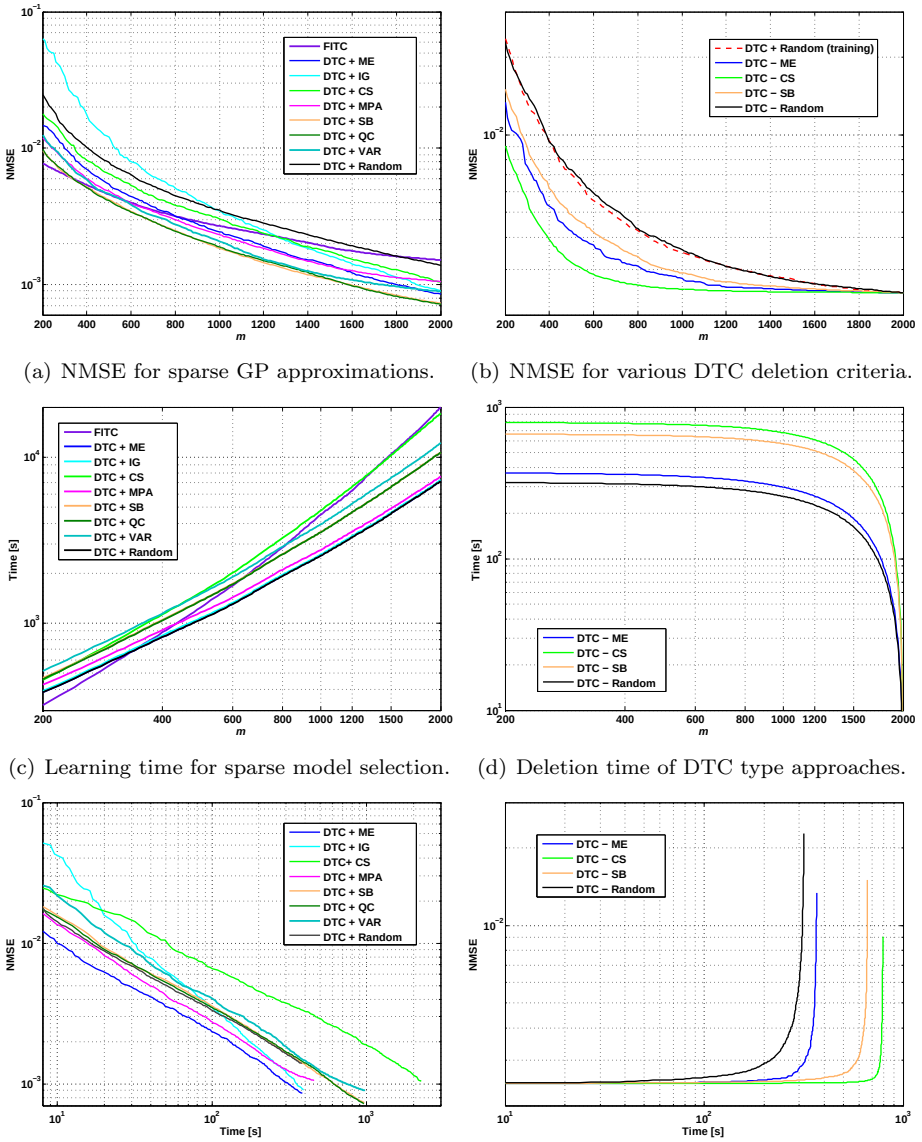
where now $\boldsymbol{\alpha}_I = \boldsymbol{R}^{-1}\boldsymbol{Q}^T\boldsymbol{K}_{I,\cdot}\boldsymbol{y} \in \mathbb{R}^m$. Note that we use the maximum error ${}_{\text{ME}}\Delta_i$ instead of the expensive projection-induced error as in [9]. Thus, we obtain the same low complexity for deletion criterion evaluation of $\mathcal{O}(1)$ per active point. Here, we coupled the error of an active training point at the current sparse model (1) with its importance under prediction. So our deletion criterion controls the current model accuracy and the generalization capability. A longer version of the paper will include more details about the entire learning process.

4 Evaluations

In this section, we compare our maximum error selection and deletion criteria against many other methods for the DTC approximation. Furthermore, we consider the FITC approximation [4] to present an extensive comparison. For all experiments we use the stationary squared exponential covariance function with automatic relevance determination, see [2]. For evaluations, we choose a benchmark data set from the SARCOS master arm (13922 training and 5569 test points), see [10]. Each point of the data set has 21 input dimensions and 7 targets, i.e. one moment for each degree of freedoms (DoF) of the robot arm.

The convergence trends with respect to the NMSE (normalized mean squared error) of many sparse GP approximations on the first DoF from the real SARCOS test data are shown in Figure 1(a). The NMSE results for randomized selection in the DTC approximation are averaged over ten runs, but for the DTC deletion schemes in Figure 1(b) we use only one random model training to demonstrate all effects caused through the criteria. The NMSE results, the complete learning times in Figure 1(c), and training times for the deletion criteria, see Figure 1(d), were captured every tenth active or virtual training points for all learning curves. The inclusion curves by Smola and Bartlett [1] and Quiñonero-Candela [8] nearly match in costs and NMSE results. The variational framework by [6] leads to constant higher effort in the learning process, e.g. compared to the curve by [8], since the regularization term increase the costs for gradient based optimization techniques. Our maximum error approach outperforms all DTC selection criteria with respect to training times for low NMSE values on test data, see Figure 1(e). For large active set sizes we nearly reach the same accuracy as the selection heuristic by [1] or [8] and outperform the matching pursuit approach by [7], see Figure 1(a). Regarding the right column in Figure 1, we outperform the DTC deletion criterion by [1] and the randomized version with respect to generalization accuracy. We also yield a good compromise between low computational effort

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.



(a) NMSE for sparse GP approximations.

(b) NMSE for various DTC deletion criteria.

(c) Learning time for sparse model selection.

(d) Deletion time of DTC type approaches.

(e) Selection time of DTC methods vs. NMSE.

(f) Deletion time of DTC methods vs. NMSE.

Figure 1: Convergence trends in NMSE and learning times on the first degree of freedom (DoF) from the real SARCOS test data for many sparse GP approximations, i.e. by Smola and Bartlett (SB) [1], by Seeger et al. (IG) [5], by Titsias (VAR) [6], by Keerthi and Chu (MPA) [7], by Quiñonero-Candela (QC) [8], and by Csató (CS) [9] (left column). The right column shows different deletion criteria of the DTC approximation. Our novel strategies (ME) give the best trade-off between low computing times and accurate prediction, i.e., see 1(e) where we yield the lowest learning curve.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

and high prediction precision as shown in Figure 1(f). All intelligent deletion schemes yield better NMSE results than the randomized deletion.

5 Conclusion

Here we proposed a very fast greedy insertion and deletion scheme for sparse GPR or, more precisely, for the DTC approximation. Our criterion is based on the maximum error between model and training data and we provided justification for this choice. It leads to a stable and efficient way for automatic sparse model selection. The primary advantage of our maximum error greedy selection is the combination of high accuracy with low computational costs for criterion calculation of all remaining points. In contrast, the insertion methods in [1, 6, 7, 8] have to select from a small random subset of remaining points for criteria evaluation. This random restriction can lead to poorer results on especially harder regression tasks. Even without caching, we are already nearly as fast as a randomized insertion. For the removal of active points our approach nearly reaches the accuracy of Csató's deletion method, outperforms the deletion criterion by [1], and is still almost as fast as a randomized removal. Compared to the FITC approximation [4], all DTC methods lead to higher prediction accuracy and lower learning times. More details about sparse GPR approximations and the relationships between the various criteria can be given in an extended version of the paper.

References

[1] A. J. Smola and P. L. Bartlett. Sparse Greedy Gaussian Process Regression. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *NIPS*, volume 13, pages 619–625, 2001.

[2] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[3] J. Quiñonero-Candela and C. E. Rasmussen. A Unifying View of Sparse Approximate Gaussian Process Regression. In R. Herbrich, editor, *JMLR*, pages 1939–1959, 2005.

[4] E. L. Snelson and Z. Ghahramani. Sparse Gaussian Processes Using Pseudo-Inputs. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *NIPS*, volume 18, pages 1257–1264, 2006.

[5] M. Seeger, C. K. I. Williams, and N. D. Lawrence. Fast Forward Selection to Speed up Sparse Gaussian Process Regression. In C. M. Bishop and B. J. Frey, editors, *AISTATS*, pages 205–212, 2003.

[6] M. K. Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In D. van Dyk and M. Welling, editors, *AISTATS*, pages 567–574, 2009.

[7] S. S. Keerthi and W. Chu. A Matching Pursuit Approach to Sparse Gaussian Process Regression. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *NIPS*, volume 18, pages 643–650, 2006.

[8] J. Quiñonero-Candela. *Learning with Uncertainty – Gaussian Processes and Relevance Vector Machines*. Phd thesis, Technical University of Denmark, 2004.

[9] L. Csató and M. Opper. Sparse Representation for Gaussian Process Regression. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *NIPS*, volume 13, pages 444–450, 2001.

[10] D. Nguyen-Tuong, J. Peters, and M. Seeger. Local Gaussian Process Regression for Real Time Online Model Learning and Control. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *NIPS*, volume 21, pages 1193–1200, 2009.