Supplementary Materials:
Sparse within Sparse Gaussian Processes using Neighbor Information

A. Details of the One-dimensional Regression Example of Section 3

For the experiment of Figure 2 of the main paper, we have executed SVGP and SWSGP in identical settings. The total inducing points we use is set to 128 which are initialized as a regular grid over one-dimensional input space. For SWSGP, the size of neighbor area is set to 16. Both of approaches use RBF kernel with variance of 0.8 and lengthscales of 0.04 and Gaussian likelihood with variance 0.01. During the training phase of SVGP and SWSGP, we fix the kernel parameters, likelihood variance and inducing inputs $\mathbf{Z}$. Thus we can see how the two methods approximate the same target posterior.

B. Further visualizations on low-dimensional data sets

We demonstrate SWSGP on one-dimensional regression problem. We have generated a synthetic data-set by sampling inputs x_i from the interval $[-2, 2]$; the targets have been computed as $y_i = \sin(12x_i) + 0.66\cos(25x_i) + \epsilon$, where ϵ is additive Gaussian noise with variance 0.1. Figure 1 summarizes the regression result for a fixed M, while the value of H varies from 4 to 64. We notice that the predictive means are nearly identical across the different sub-figures. These observations suggest that SWSGP is able to work and converge well even though H is significantly less than M.

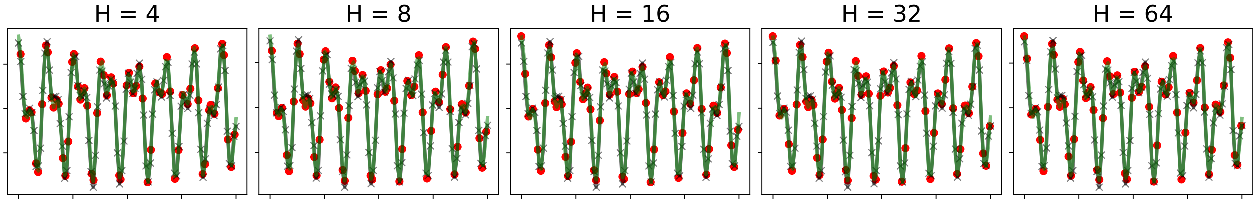

Figure 1: SWSGP is applied on a one-dimensional data set, where M is fixed to 64 and H is increased gradually from 4 to 64. The red dots are inducing positions; the black crosses are testing points; the green line refers to predictive means.

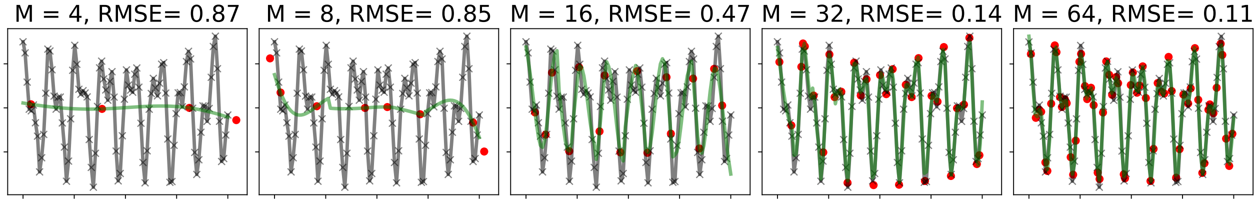

Figure 2: SWSGP is applied on 1D. The red dots are inducing positions. The black crosses are testing samples. The green lines are predictive means. The title of each sub-figures shows M and corresponding RMSE.

We also show that the performance of SWSGP improves when increasing the total number of inducing points while keeping the number of active inducing points H fixed. We intuitively expect that a larger the total number of inducing points should translate to a more accurate model. In these experiments, the size of neighbor area is fixed to 4, i.e. $H = 4$, and the total number of inducing points are varies from 4 to 64. We see that the sequence of the predictive means in Figure 2 are more and more accurate from left to right. Although we are using a small neighbor area, our model is improved when increasing the total number of inducing points.

In Figure 3, we examine SWSGP on a two-dimensional classification data set (BANANA), where M is fixed to 64 and H is

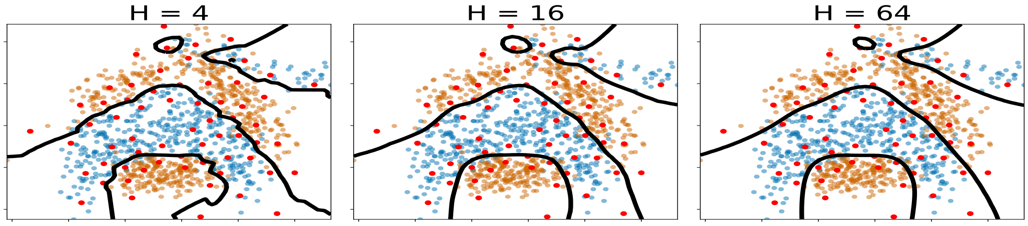

Figure 3: Visualization of SWSGP on BANANA data sets with increasing H. The total number of inducing points M is fixed to 64, while the size of neighbor area H varies from 4 to 64. The red dots represent the inducing inputs. The orange and blue dots are training points from two different classes. The black lines are the contours of a classifier where the predictive mean is 0.5.

increased from 4 to 64. In general, these boundaries remain sensible across the whole range of values of H, suggesting that SWSGP is able to work and converge well even though H is significantly less than M. We also observe that the contours of the classifier become smoother as H is increasing.

We also illustrate that these contours become more sensible, where H is fixed to 4 and M is gradually increased from 4 to 64. In Figure (4) we see that the classification boundaries improve when increasing M, even though the number of active neighbors is fixed.

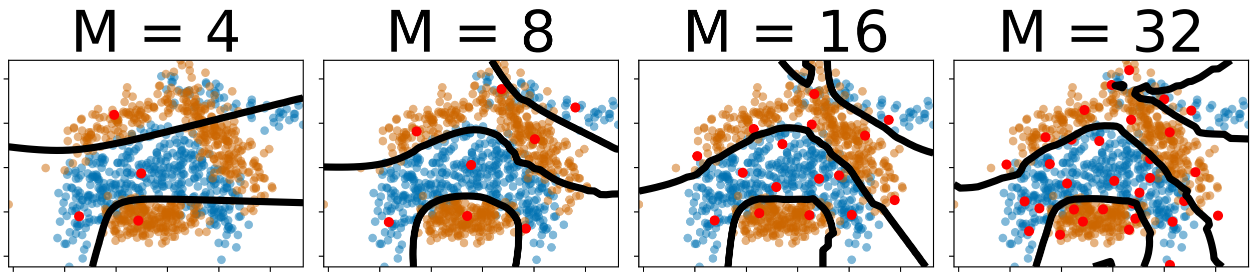

Figure 4: Visualization of SWSGP on BANANA data sets with increasing M. The size of neighbor area H is set to 4. The total number of inducing points M varies from 4 to 64. The red dots represent inducing inputs. The orange and blue dots are the input points from the two different classes. The black lines are the contours of a classifier where the predictive mean is 0.5.

C. Various Options for H-nearest Inducing Points Selection

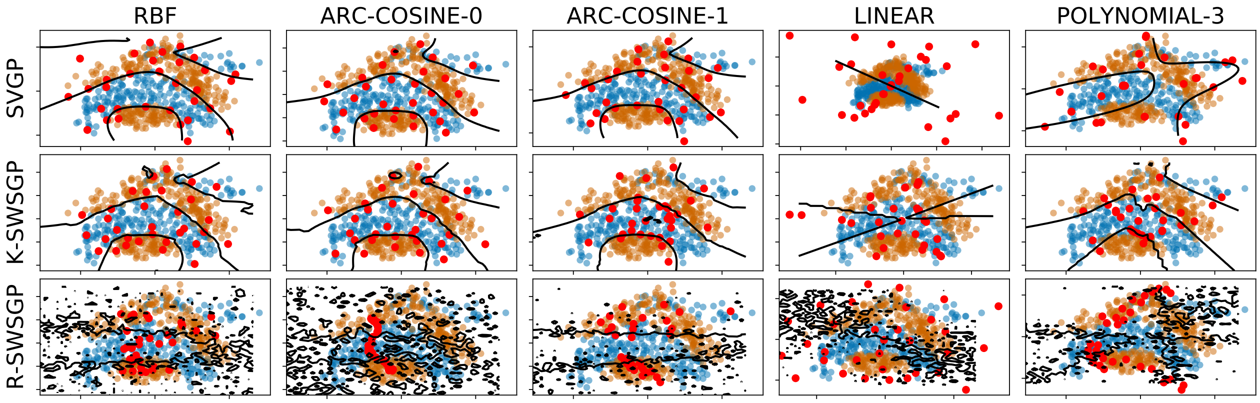

Figure 5: SWSGP on various kernels and strategies for selecting the H-nearest inducing points.

As we discuss in the paper, the selection of H-nearest inducing points $\mathbf{Z}_x^H$ is made by using the kernel as a proxy to the concept of distance. Intuitively, a kernel defines the similarity between two points in the input space, which is more formally expressed as correlation. The kernel implicitly defines a kind of distance that we use to determine the active neiborhood. Thus, the selected neiborhood is dominated by the inducing points with largest kernel values.

In the main paper, we have used different versions of the Matérn kernel. We shall now explore the effect of our neiborhood-selection strategy on a number of different kernels, both stationary and non-stationary. We apply SWSGP on the BANANA data-set using different heuristics for the H-nearest inducing points selection. Let K-SWSGP denote what is essentially the vanilla version of our method, where the kernel-based heuristic is used as a proxy to distance. In the case of the RBF kernel, K-SWSGP essentially corresponds to the Euclidean distance. We also examine a random-based heuristic (R-SWSGP) in which H-nearest inducing points are randomly chosen. In all cases, we set M and H as 32 and 8 respectively. We also compare against SVGP with M of 32.

In Figure (5), we visualize the contours of classifiers of SVGP and SWSGP with various configurations. Clearly, R-SWSGP does not work, i.e. the contours are discontinuous and the locations of contours does not makes sense. Regarding the kernels RBF, ARC-COSINE-0 and ARC-COSINE-1, our method (K-SWSGP) seems to be virtually identical to SVGP. The advantages of K-SWSGP over SVGP are shown when using POLYNOMIAL-3. It is highly possible that the flexibility of variational distribution over inducing variables, i.e. $q(\mathbf{u})$, in SWSGP is the main reason for this difference.

D. Corner Cases

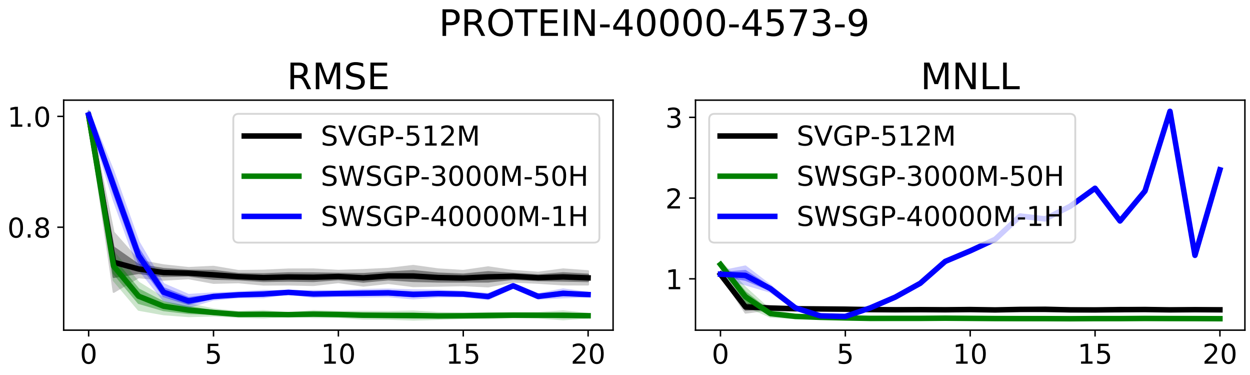

Figure 6: 'Corner' case of SWSGP. The Horizontal axis indicates the training time (in minutes).

We investigate a corner case of SWSGP where the set of inducing points are fixed to the training point, i.e. $\mathbf{Z} = \mathbf{X}$, and the approximated posteriors of each training point within mini-batches are computed by considering only one nearest inducing point, i.e. $H = 1$. Intuitively speaking, the prediction at an unseen data point $\mathbf{x}_*$ is made by using a GPs with only one observation that is the nearest point to $\mathbf{x}_*$. Figure 6 indicates that the corner case where $\mathbf{Z}$ is set by $\mathbf{X}$ and H is equal to 1 is prone to overfitting. This can be shown by the growth of MNLL over time. We speculate that the ratio of H to M that is too small leads to the loss of essential information to produce good predictions.