Active Learning with Maximum Margin Sparse Gaussian Processes

Weishi Shi **Qi Yu**
Rochester Institute of Technology
{ws7586, qi.yu}@rit.edu

Abstract

We present a maximum-margin sparse Gaussian Process (MM-SGP) for active learning (AL) of classification models for multi-class problems. The proposed model makes novel extensions to a GP by integrating maximum-margin constraints into its learning process, aiming to further improve its predictive power while keeping its inherent capability for uncertainty quantification. The MM constraints ensure small "effective size" of the model, which allows MM-SGP to provide good predictive performance by using limited "active" data samples, a critical property for AL. Furthermore, as a Gaussian process model, MM-SGP will output both the predicted class distribution and the predictive variance, both of which are essential for defining a sampling function effective to improve the decision boundaries of a large number of classes simultaneously. Finally, the sparse nature of MM-SGP ensures that it can be efficiently trained by solving a low-rank convex dual problem. Experiment results on both synthetic and real-world datasets show the effectiveness and efficiency of the proposed AL model.

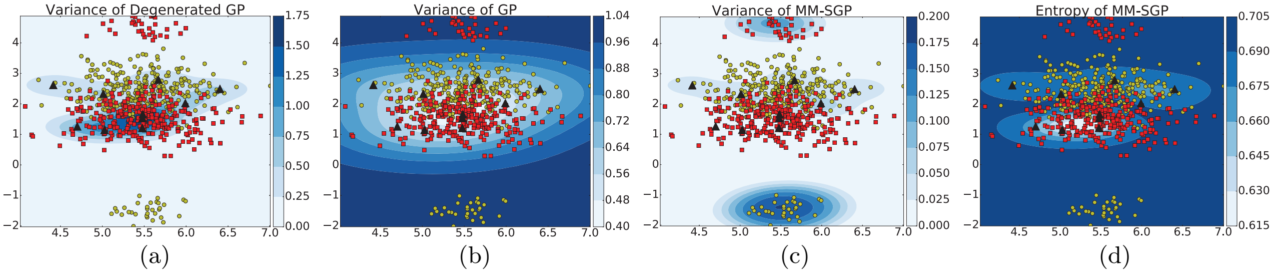
1 Introduction

Active learning (AL) provides an effective means to reduce human labeling effort by selecting the most informative data samples for more effective model training [1]. Previous work on AL has been successfully applied to various applications with promising results [2, 3, 4]. In particular, an effective AL model should possess three key properties: i) good predictive performance upon convergence, ii) accurate uncertainty estimation for effective active sampling, and iii) efficient model training to support real-time human-machine collaboration. We propose a *maximum-margin sparse Gaussian process (MM-SGP)* for active learning of multi-class classification models, which simultaneously meets these key properties. By leveraging a flexible covariance function, a GP naturally captures the correlations of data samples to achieve good prediction performance while offering important uncertainty information that is critical for active sampling. We make novel extensions to a GP by integrating maximum-margin (MM) constraints into its learning process to further improve its predictive power. Furthermore, the MM constraints allow the MM-SGP to inherit key properties of other maximum-margin models (e.g., SVMs) that can make good predictions by only relying on limited data samples. Such behavior indicates the great potential of using MM-SGP for AL, which aims to choose these most useful data samples for model training through effective sampling.

In particular, the MM constraints are integrated using the maximum entropy discrimination (MED) framework [5, 6], which learns a parameter distribution with minimum assumptions among all feasible choices (as indicated by 'maximum entropy' in the name). As the MED framework directly works on model parameters (i.e., weights), we leverage the 'parameter-space' view of a GP to make the model parameters explicit by placing a proper prior distribution to ensure equivalence to its 'function-space' view. This transformation elegantly brings a nonparametric model (GP) and a parametric framework (MED) together under a unified learning scheme. As a result, the proposed MM-SGP further enhances the generalization capacity of a standard GP, which is clearly demonstrated by its superior predictive accuracy as evidenced by our experiments over both synthetic and real-world data.

Since the MM constraints are applied to the model parameters, these parameters need to be made explicit in the MM-SGP model. While directly using the learned parameters still provide the same (mean) predictions, it will *degenerate* the GP that can negatively impact

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).



(a) (b) (c) (d)

Figure 1: Predictive variance of (a) degenerated GP, (b) GP, and (c) MM-SGP; (d) entropy of MM-SGP predicted class labels over a dataset with two classes (similar to the predicted entropy of a standard GP, which not shown here), each of which comes from a mixture of two Gaussian's, where labeled samples are shown in black triangles.

data sampling in AL. In particular, a degenerate GP tends to underestimate the predictive variance of unlabeled data samples that are far away from the current training data [7], which will result in the sampling bias issue that causes slow and/or inaccurate convergence of AL [8]. Figure 1(a) demonstrates the undesired predictive variance provided by a degenerate GP trained over a dataset with two classes (shown as yellow and red), each of which is a mixture of two Gaussian's. The black triangles represent the labeled data and the squares/circles are unlabeled with colors indicating their (unknown) class labels. Since the model is trained using limited labeled data (typical for AL) from the two major clusters in the center, data samples from the two faraway minor clusters are predicted with very low variance. If the predictive variance is used as an uncertainty measure, it will mislead active sampling and suggest the AL model not to choose important data samples faraway from the current training data. This may lead to a wrong classification boundary that completely misclassify the two minor clusters.

Fortunately, the undesirable predictive variance can be systematically addressed through augmentation that assigns an additional weight parameter to each test data sample [9]. Augmentation has been successfully applied to several GP based finite linear models, including Subset of Regressors (SoR) and Relevant Vector Machines (RVM), which suffer from underestimated predictive variances [7]. However, augmenting each test sample in a large unlabeled pool for active sampling is computational prohibitive, making it infeasible to support real-time human-machine interactions in AL. We propose to identify a set of most representative unlabeled data samples that cover critical areas in the sample space by performing a *one-time joint augmentation* with much improved sampling efficiency. Figure 1(c) shows that the proposed joint augmentation accurately estimates the predictive variances of data samples from the two minor clusters, which will allow them to be properly sampled in the early stage of AL so that a correct decision boundary can be formed to improve convergence.

The predictive variance provides important information that complements the uncertainty of the predicted class distribution, captured by class entropy. Figure 1(d) shows that a high class entropy is assigned to both data samples close to or faraway from the training data. However, these data samples contribute differently by refining the current decision boundary or exploring critical areas in the sample space, respectively. We will develop a novel sampling function that *aggregates both the predicted class entropy and the predictive variance* of MM-SGP.

Finally, as the model parameters need to be continuously learned along with the AL process, we propose to use a sparse GP that chooses only a selected subset of support points (or pseudo inputs) to approximate the entire training set. By joint augmentation of the support points, each augmented unlabeled sample has direct access to all the labeled training samples instead of through the sparse support points. This ensures that the predictive variances of data samples close to the training data are not overestimated as sparsity level increases. The sparse nature of MM-SGP ensures that it can be efficiently trained by solving a *low-rank convex dual problem* for fast model training that is 3-5X more efficient than its dense version. In fact, a dense GP starts to pose a computational bottleneck in middle and later phases of AL when more labeled data is accumulated as observed in our experiments on real-world data. Our key contributions include: (1) a unified learning scheme that seamlessly integrates the maximum-margin constraints with a nonparametric GP, (2) a novel sampling function that combines entropy and predictive variance, and (3) leveraging the low-rank structure of a sparse GP for efficient and scalable model training to support realtime AL.

2 Related Work

GP has been used as popular statistical learning tool for active data selection in regression problems [10, 11, 12, 13]. To extend a GP to classification, sigmoid or softmax transformation of the latent Gaussian variables is needed, which makes the exact inference infeasible and poses a challenge for AL. To avoid this extra complexity, the ratio between the posterior mean and variance of the latent Gaussian variables has been used for data sampling [14]. However, the Gaussian variables do not transform linearly through the sigmoid/softmax functions, leading to inaccurate uncertainty estimation. Through approximate inference, the predictive distribution over class labels can be computed from a GP and used for active sampling [15]. However, the predicted label distribution only provides the mean prediction without predictive variance, which does not provide an effective way to explore the unlabeled data space for data sampling. Furthermore, the parametric version of GP is adopted [15], which is similar to the relevant vector machine (RVM) based active learners [16]. RVM has also been combined with an SVM, leading to a kernel committee machine (KMC) that is suitable for active learning [17]. However, due to the degenerate nature, the parametric GP models (e.g., SoR and RVM) tend to predict low variance for samples faraway from current training data [18], which is undesirable to explore the unlabeled data space. Our experimental results confirm the advantage of the proposed MM-SGP over these methods.

As a maximum-margin model, SVMs have been commonly used for AL using the distance to the decision boundary to quantify uncertainty for sampling [19, 20]. For multi-class problems, due to the interplay of multiple decision surfaces, the uncertainty of a data sample cannot be easily characterized by distances [21]. A Best-versus-Second Best (BvSB) model was developed to choose the samples with the closest top-two probable classes [22]. However, BvSB ignores the probability distribution of other classes, making it less effective for many-class settings. Entropy has also been used to quantify uncertainty for multi-class AL [23, 24]. The challenge lies in how to accurately estimate entropy especially during the early phase of AL. The estimated decision-boundary can be highly inaccurate (as shown in our experiments) especially when a large number of classes are involved.

Probabilistic models provide an alternative way to consider all potential classes. For example, the expected error within a neighbourhood of a candidate data sample has been used as the sampling score [25]. However, the high computational cost makes it infeasible to scale to a large number of classes. A convex hull-based sampling function is used to choose data samples with the potential to significantly change the current model [26]. However, the model solely relies on an SVM and its support vectors for data sampling, which may be limited in exploring the data space.

To further advance frontiers in AL, the proposed MM-SGP integrates a GP with the MM constraints through a unified learning scheme so that it benefits from both the *exploring capacity* of the former and the *discriminative power* of the latter for most effective AL.

3 The MM-SGP Model

We denote a training set with N data samples as $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ and let $\mathbf{y} = \{y_1, ..., y_N\}$ be their corresponding labels. Consider the binary-class case where $\forall y_n \in \mathbf{y}, y_n \in \{-1, +1\}$ and the multi-class problems can be achieved via one-versus-the-rest. The conditional distribution of label y_n is given by $p(y_n = +1|f(\mathbf{x}_n)) = \sigma(f(\mathbf{x}_n))$, where σ is the logistic sigmoid function and $f(\mathbf{x}_n)$ is a latent function introduced by GP, which models the log odd of sample $\mathbf{x}_n$ assigned label y_n. A prior distribution is further placed on function $f(\mathbf{x}_n)$, given by $p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, K)$, where $\mathbf{f} = (f_1, ..., f_N)^T$ with $f_n = f(\mathbf{x}_n)$. K is the covariance matrix with $K_{nm} = k(\mathbf{x}_n, \mathbf{x}_m)$, where $k(\cdot, \cdot)$ is a kernel function. Prediction on a test sample $\mathbf{x}_*$ can be achieved in two steps: (1) computing the predictive distribution of f_* : $p(f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \int p(f_*|\mathbf{X}, \mathbf{x}_*, \mathbf{f})p(\mathbf{f}|\mathbf{X}, \mathbf{y})\mathrm{d}\mathbf{f}$, where the posterior is given by $p(\mathbf{f}|\mathbf{X}, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})$, and (2) making a prediction by integrating out the predictive distribution: $p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \int p(y_*|f_*)p(f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)\mathrm{d}f_*$. The challenge lies in the likelihood term $p(y_n|f_n)$, which is a logistic function and non-conjugate to the Gaussian prior. Thus, further approximation is needed to compute the integration, which will be detailed later.

In order to incorporate the maximum-margin (MM) constraints that directly operate on the model parameters (i.e., weights), we transform the GP model described above into its weight-space representation. In particular, let $\mathbf{f} = K\mathbf{w}$ and by placing a prior $p(\mathbf{w}|\mathbf{X}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, K^{-1})$, we recover $p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, K)$. In this weight-space view, each data sample $\mathbf{x}_n$ is essentially represented by a feature vector $\phi(\mathbf{x}_n) = \mathbf{k}_{\mathbf{x}_n}$, where $[\mathbf{k}_{\mathbf{x}_n}]_j = k(\mathbf{x}_n, \mathbf{x}_j)$. To achieve fast training of the GP for real-time AL, we further propose to use a sparse version of GP, which constructs a set of support points $\overline{\mathbf{X}} = \{\overline{\mathbf{x}}_m\}_{m=1}^M$, where $M < N$, for model training and prediction. Under a sparse GP, we have $\mathbf{f} = K_{NM}\mathbf{w}_M$ with a prior $p(\mathbf{w}_M|\overline{\mathbf{X}}) = \mathcal{N}(\mathbf{w}_M|\mathbf{0}, K_{MM}^{-1})$, where $[K_{NM}]_{nm} = k(\mathbf{x}_n, \overline{\mathbf{x}}_m)$, $[K_{MM}]_{mm'} = k(\overline{\mathbf{x}}_m, \overline{\mathbf{x}}_{m'})$, and $\overline{\mathbf{x}}_m \in \overline{\mathbf{X}}$. The construction of $\overline{\mathbf{X}}$ is covered in Sect. 3.

The MED framework aims to find an optimal parameter distribution $q(\mathbf{w})$ that minimizes a regularized KL-divergence: $KL(q(\mathbf{w})||p(\mathbf{w}_0)) + U(\xi)$, where $p(\mathbf{w}_0)$ is a prior distribution and $U(\xi)$ is a function over the slack variables introduced along with the MM constraints given by $y_n(\mathbf{w}^T\mathbf{k}_{\mathbf{x}_n}) \geq \Delta l_n(y) - \xi_n$ with $\xi_n \geq 0$ and $\Delta l_n(y)$ being a loss function (e.g., 0-1 loss). Intuitively, an optimal $q(\mathbf{w})$ should make minimum additional assumptions beyond the given prior $p(\mathbf{w}_0)$ (i.e., "maximum entropy" in MED) besides meeting the MM constraints. In addition to replacing the prior distribution $p(\mathbf{w}_0)$ with $p(\mathbf{w}_M|\overline{\mathbf{X}})$, we make another key extension to the MED framework by further incorporating the expected log likelihood $\mathbb{E}_{q(\mathbf{w})}[\ln p(\mathbf{y}|\mathbf{X},\overline{\mathbf{X}},\mathbf{w})]$ into the objective function. In this way, the optimal $q(\mathbf{w})$ essentially minimizes the negative lower bound (hence maximizes the lower bound) of the log marginal likelihood while meeting the MM constraints. To see this, by applying Jensen's inequality to the log marginal likelihood, we obtain a lower bound given by $\ln p(\mathbf{y}|\mathbf{X},\overline{\mathbf{X}}) \geq \mathbb{E}_{q(\mathbf{w})}[\ln p(\mathbf{y}|\mathbf{X},\overline{\mathbf{X}},\mathbf{w})] - KL(q(\mathbf{w})||p(\mathbf{w}|\overline{\mathbf{X}}))$. Adding the MM constraints gives

$$\begin{aligned} \min_{q(\mathbf{w}),\boldsymbol{\gamma},\overline{\mathbf{X}},\boldsymbol{\xi}} \; & c\sum_n \xi_n + KL(q||p) - \mathbb{E}_{q(\mathbf{w})}[\ln g(\mathbf{w},\boldsymbol{\gamma},\overline{\mathbf{X}})] \\ \text{s.t.} \quad & \forall n: \quad \mathbb{E}_{q(\mathbf{w})}[y_n(\mathbf{w}^T\mathbf{k}_{\mathbf{x}_n})] \geq \Delta l_n(y) - \xi_n, \\ & \xi_n \geq 0, \quad \int q(\mathbf{w})\mathrm{d}\mathbf{w} = 1 \end{aligned} \tag{1}$$

where $KL(q||p) = KL(q(\mathbf{w})||p(\mathbf{w}|\overline{\mathbf{X}}))$, $[\mathbf{k}_{\mathbf{x}_n}]_m = k(\mathbf{x}_n, \overline{\mathbf{x}}_m), \forall \overline{\mathbf{x}}_m \in \overline{\mathbf{X}}$, and $\boldsymbol{\xi} = (\xi_1, ..., \xi_N)^T$. One additional change we make is to replace the likelihood $p(\mathbf{y}|\mathbf{X},\overline{\mathbf{X}},\mathbf{w})$ by an exponential quadratic lower bound function $g(\mathbf{w},\boldsymbol{\gamma},\overline{\mathbf{X}})$, given by

$$\begin{aligned} g(\mathbf{w},\boldsymbol{\gamma},\overline{\mathbf{X}}) = \prod_{n=1}^{N} \sigma(\gamma_n)\exp\{ & \frac{y_n(\mathbf{w}^T\mathbf{k}_{\mathbf{x}_n}) - \gamma_n}{2} - \\ & \lambda(\gamma_n)([\mathbf{w}^T\mathbf{k}_{\mathbf{x}_n}]^2 - \gamma_n^2)\} \end{aligned} \tag{2}$$

where $\lambda(\gamma) = \frac{1}{2\gamma}(\sigma(\gamma) - \frac{1}{2})$ [27]. This is because the non-conjugate logistic function in the expected likelihood does not lead to an analytical form for $q(\mathbf{w})$, which makes the original optimization problem difficult to solve. By using the lower bound, $q(\mathbf{w})$ will follow a Gaussian distribution governed by local variational parameters $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_N)^T$ as shown in Sect. 3.

Since (1) is convex over $q(\mathbf{w})$ with linear constraints, dual sparsity is ensured from the KKT conditions of the Lagrangian function of (1). Thus, only a subset of Lagrangian multipliers will be non-zero that correspond to the active MM constraints. This key property allows MM-SGP to provide good predictive performance by using limited "active" data samples, similar to other maximum-margin models. Meanwhile, the small effective size of MM-SGP clearly shows its potential as an AL model. Through effective sampling (Sect. 3.1), we aim to choose labeling only these "active" data samples that play an effective role in the model's prediction power. To extend to K classes, we adopt one-versus-the-rest and then apply a softmax transformation, which gives rise to the posterior probability of the k-th class: $p(y = C_k|\mathbf{x}) = \mathbb{E}_{q(W)}[e^{\mathbf{w}^{(k)^T}\mathbf{k}_\mathbf{x}} / \sum_{j=1}^K e^{\mathbf{w}^{(j)^T}\mathbf{k}_\mathbf{x}}]$, where $q(W) = \prod_k q(\mathbf{w}^{(k)})$. Since the expectation cannot be computed analytically, we perform Monte Carlo (MC) integration by drawing samples from $q(W)$, softmax them, and then average.

Posterior Inference and Parameter Learning

In this section, we present a principled optimization framework that leverages the convexity of (1) over the variational distribution $q(\mathbf{w})$ and the reduced rank of the MM-SGP for efficient posterior inference and parameter learning. Overall, the framework will adopt an iterative process to update the variational distribution $q(\mathbf{w})$ (along with $\boldsymbol{\xi}$) and other key model parameters $(\boldsymbol{\gamma}, \overline{\mathbf{X}})$ in a coordinate decent fashion until convergence. We present the key results in this section and leave the detailed proofs to Appendices A and B.

Posterior Inference: First, by recognizing that (1) is a convex problem over the variational distribution $q(\mathbf{w})$, we introduce Lagrangian multipliers α_n's and α_0 for each inequality and equality constraints, respectively, which gives the Lagrangian function $L(q(\mathbf{w}), \boldsymbol{\xi}, \overline{\mathbf{X}}, \boldsymbol{\gamma}, \boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_0, ..., \alpha_N)^T$. We start by solving $q(\mathbf{w})$ while fixing other parameters. This can proceed by taking the partial derivative over $q(\mathbf{w})$ using the calculus of variations to the Lagrangian function to obtain an optimal analytical form dependent on the Lagrangian multipliers. We summarize our major results below.

Lemma 1 *The optimal $q(\mathbf{w})$ follows a Gaussian distribution: $\hat{q}(\mathbf{w}) = \mathcal{N}(\boldsymbol{\mu}_q(\boldsymbol{\alpha}), \Sigma_q)$, where*

$$\boldsymbol{\mu}_q(\boldsymbol{\alpha}) = \Sigma_q[\sum_{n=1}^{N}((\alpha_n + \frac{1}{2})y_n\mathbf{k}_{\mathbf{x}_n}], \tag{3}$$

$$\Sigma_q^{-1} = K_{MM} + 2\sum_{n=1}^{N}\lambda(\gamma_n)\mathbf{k}_{\mathbf{x}_n}\mathbf{k}_{\mathbf{x}_n}^T \tag{4}$$

It is worth to note that the optimal parameter distribution $\hat{q}(\mathbf{w})$ is closely related to the normal equations. In particular, substituting Σ_q^{-1} to $\boldsymbol{\mu}_q$, we have

$$\boldsymbol{\mu}_q(\boldsymbol{\alpha}) = (2K_{NM}^T\Gamma K_{NM} + K_{MM})^{-1}(K_{NM}^T \operatorname{diag}(\widehat{\boldsymbol{\alpha}})\mathbf{y}) \tag{5}$$

The self-projection $K_{NM}^T K_{NM}$ and target projection $K_{NM}^T\mathbf{y}$ in normal equations are weighted by local variational parameters $\Gamma = \operatorname{diag}(\lambda(\gamma_1), .., \lambda(\gamma_N))$ and Lagrangian multipliers $\widehat{\boldsymbol{\alpha}} = (\alpha_1 + \frac{1}{2}, .., \alpha_N + \frac{1}{2})^T$, respectively. As a result, *the important data samples and responses will be assigned a higher weights through their corresponding λ's and α's.* The first term on the r.h.s. is further regularized by the precision matrix of the SGP prior.

By substituting $\hat{q}(\mathbf{w})$ back into L, we obtain the Lagrangian dual of (1), which takes the form of a constrained quadratic programming (QP) problem with a low-rank structure.

Theorem 1 *The Lagrangian dual of the primal problem in* (1) *takes the form of a low-rank QP:*

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T Q \boldsymbol{\alpha} \quad s.t. \sum_{n=1}^{N} \alpha_n = c, \alpha_n \in [0, \infty) \tag{6}$$

where Q has a low-rank representation as $Q = VV^T$ with $V = \text{diag}(\mathbf{y}) K_{NM} \Sigma_q^{\frac{1}{2}}$.

It is straightforward to see from the above theorem that $\text{rank}(Q) = M < N$ given that $\text{rank}(V) = M$. The bound of time complexity of each step in the iterative QP solving process is reduced to $O(NM^2)$ from $O(N^3)$. Since solving the dual problem is the computationally most expensive component of MM-SGP, leveraging the special low-rank structure ensures that the dual problem can be solved much more efficiently [28, 29], as evidenced by our experiments.

Parameter Learning: After solving the Lagrangian multipliers, an optimal $\hat{q}(\mathbf{w})$ is obtained. We substitute it back to L, which turns it into a function of only $\boldsymbol{\gamma}$ and $\overline{\mathbf{X}}$:

$$\begin{aligned} L(\overline{\mathbf{X}}, \boldsymbol{\gamma}) =& KL(q||p) - \mathbb{E}_{q(\mathbf{w})}[\log g(\mathbf{w}, \boldsymbol{\gamma}, \overline{\mathbf{X}})] \\ &- \sum_n \alpha_n \left\{ \mathbb{E}_{q(\mathbf{w})}[y_n(\mathbf{w}^T \mathbf{k}_{\mathbf{x}_n})\right\} + \text{const} \end{aligned} \tag{7}$$

where terms do not involve $\boldsymbol{\gamma}$ and $\overline{\mathbf{X}}$ are absorbed into the 'const' term. To derive the closed-form update rule for $\boldsymbol{\gamma}$, we set $\frac{\partial L}{\partial \boldsymbol{\gamma}} = 0$ and get

$$\gamma_n^2 = \mathbf{k}_{\mathbf{x}_n}^T \mathbb{E}[\mathbf{w}\mathbf{w}^T] \mathbf{k}_{\mathbf{x}_n} = \mathbf{k}_{\mathbf{x}_n}^T (\Sigma_q + \mu_q(\boldsymbol{\alpha})\mu_q(\boldsymbol{\alpha})^T) \mathbf{k}_{\mathbf{x}_n} \tag{8}$$

Next, we show the learning of the hyperparameters of the kernel function. Let θ denote a hyperparameter (e.g., characteristic length scale of a RBF kernel), which contributes to $L(\overline{\mathbf{X}}, \boldsymbol{\gamma})$ through two kernel matrices K_{NM} and K_{MM}. We use $F(\overline{\mathbf{X}})$ and $G(\overline{\mathbf{X}})$ to denote these two matrices to make their dependencies of $\overline{\mathbf{X}}$ (also also θ) explicit. This will also facilitate our derivation of the update rule of $\overline{\mathbf{X}}$. We first reformulate the moments of $q(\mathbf{w})$ as:

$$\mu_q(\boldsymbol{\alpha}) = \Sigma_q F(\overline{\mathbf{X}})^T \Lambda \widehat{\boldsymbol{\alpha}}, \tag{9}$$

$$\Sigma_q^{-1} = G(\overline{\mathbf{X}}) + 2F(\overline{\mathbf{X}})^T \Gamma F(\overline{\mathbf{X}}) \tag{10}$$

where $\Lambda = \text{diag}(\mathbf{y})$. Since all three terms in (7) are expectations over $q(\mathbf{w})$, they can be expressed using the moments given above. Following the chain rule and using the following results, we can compute their derivatives over θ:

$$\frac{\partial \Sigma_q}{\partial \theta} = \frac{\partial G(\overline{\mathbf{X}})}{\partial \theta} + 2F(\overline{\mathbf{X}})(\Gamma + \Gamma^T)\frac{\partial F(\overline{\mathbf{X}})}{\partial \theta}, \tag{11}$$

$$\frac{\partial \mu_q(\boldsymbol{\alpha})}{\partial \theta} = \left(\frac{\Sigma_q}{\theta} F(\overline{\mathbf{X}})^T + \Sigma_q \frac{\partial F(\overline{\mathbf{X}})}{\partial \theta} \right) \Lambda \widehat{\boldsymbol{\alpha}} \tag{12}$$

Putting them together, we can perform gradient decent to optimize hyperparameter θ. The set of support points $\overline{\mathbf{X}}$ can be learned in the same way but the derivation is more involved that requires derivatives over matrices (details are provided in Appendix B.2).

3.1 MM-SGP based Active Sampling

We propose a novel AL process that integrates both the predicted class entropy and predictive variance output by the MM-SGP model for effective active sampling. Given a test data sample $\mathbf{x}_*$ and K classes, a probabilistic classifier outputs K-class probabilities: $p(y_*|\mathbf{x}_*)$, which is a K-dimensional vector $\boldsymbol{\pi}$ with $\pi_k = p(y_* = C_k|\mathbf{x}_*)$. While $p(y_*|\mathbf{x}_*)$ captures the uncertainty of the prediction over the K classes (i.e., among $\pi_k' s$), it is only a point estimate of $\boldsymbol{\pi}$ (one π_k for each class). As a Bayesian model, the proposed MM-SGP allows us to quantify the variation of each π_k using its predictive variance. The key idea is to leverage the predictive distribution of the latent function $f_*^{(k)}$ for each class, given by $p(f_*^{(k)}|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) \approx \int p(f_*^{(k)}|\mathbf{X}, \mathbf{x}_*, \mathbf{w}^{(k)}) q(\mathbf{w}^{(k)}) d\mathbf{w}^{(k)}$. Since the variational distribution $q(\mathbf{w}^{(k)})$ is a Gaussian, $p(f_*^{(k)}|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$ will also be a Gaussian, whose variance can be analytically computed. However, the final class prediction still requires a sigmoid transformation, making the computation of the final predictive variance intractable. We propose to use MC integration by drawing samples from $p(f_*^{(k)}|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$, performing sigmoid transformation, and then averaging to compute the predictive variance $\text{Var}_*^{(k)}$ of sample $\mathbf{x}_*$ for each of the K class.

By assigning a low variance to data samples near to the training data and a high variance to faraway samples, the *predictive variance complements the class entropy*, which allows the proposed sampling function to differentiate data samples based on their distinct contributions to model training and sample them accordingly. As nearby data samples are more effective to fine-tune the current decision boundary and faraway ones help better explore the data space, we propose the following function for many-class sampling:

$$\hat{\mathbf{x}}_* = \arg\max_{\mathbf{x}_*} \mathcal{H}(y_*|\mathbf{x}_*) + \eta \sum_k \text{Var}_*^{(k)}/K \tag{13}$$

where η is used to balance between class entropy and total predictive variance averaged over K classes, aiming to choose a data sample that benefits a large number of classes when K is large. Parameter η is dynamically updated to give a higher weight in the early stage of AL to the variance term for better exploration and then shift the focus to the entropy term for effective fine-tuning of decision boundaries with a correct shape obtained through effective exploration. It is worth to note that the variance term decreases fast as the AL process effectively explores the sample space so the balance between these two terms are automatically adjusted and do not rely too much on η.

However, directly applying MM-SGP may underestimate the predictive variance, especially for data samples far away from the training data. This will introduce undesirable sampling behavior. In essence, by making the model parameters explicit due to the integration of MM constraints, the GP sub-component of MM-SGP resembles the SoR algorithm [30], which suffers from a similar issue. A principled way to address this issue is through augmentation that assigns one additional weight parameter to each test data sample [9]. We summarize the key result via the following lemma [7]:

Lemma 2 *By adding an extra weight w_* to the test data sample $\mathbf{x}_*$, the augmented predictive variance of its latent function is given by*

$$\begin{aligned} v_*(\mathbf{x}_*) =& k(\mathbf{x}_*, \mathbf{x}_*) \\ & - \mathbf{k}_*^T [K_{NM} K_{MM}^{-1} K_{NM}^T + \mathbf{v}_* \mathbf{v}_*^T / c_*]^{-1} \mathbf{k}_* \end{aligned} \tag{14}$$

where $[\mathbf{k}_*]_n = k(\mathbf{x}_*, \mathbf{x}_n), \forall \mathbf{x}_n \in \mathbf{X}, \mathbf{v}_* = \mathbf{k}_* - K_{NM} K_{MM}^{-1} \mathbf{k}_{\mathbf{x}_*}$ with $[\mathbf{k}_{\mathbf{x}_*}]_m = k(\mathbf{x}_*, \bar{\mathbf{x}}_m), \forall \bar{\mathbf{x}}_m \in \overline{\mathbf{X}}$, and $c_* = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_{\mathbf{x}_*}^T K_{MM}^{-1} \mathbf{k}_{\mathbf{x}_*}$.

It is clear that the augmented predictive variance will be high if $\mathbf{x}_*$ is far from the training data as the second term in (14) will be small. Furthermore, due to augmentation, data sample $\mathbf{x}_*$ directly interacts with all the training samples (through the $\mathbf{k}_*$ term). Hence, it does not overestimate the variance if $\mathbf{x}_*$ is close to the training data. As a result, augmentation provides a *more accurate predictive variance* than other sparse GP models (e.g., Deterministic Training Conditional (DTC) model and Fully Independent Training Conditional (FITC) model) that tend to overestimate the predictive variance of nearby samples due to increased sparsity [9]. The improved variance prediction will further benefit the MM-SGP based active sampling.

Augmenting each testing sample is computational expensive ($O(NM)$ to evaluate (14)) especially for a large unlabeled candidate pool. We propose to expedite this process by performing a one-time joint augmentation of a set of representative samples from the unlabeled pool. Kernel density estimation (KDE) is used to identify data samples from densely distributed areas while being far from the training data (see Appendix C for details). Given an augmentation set $\hat{\mathbf{A}} = \{\hat{\mathbf{x}}_i\}_{i=1}^S$, we first compute an augmented posterior variance $\Sigma_q^{(k)^+}$, which is defined over $\overline{\mathbf{X}} \cup \hat{\mathbf{A}}$. Then, for a test data sample, $\mathbf{x}_*$, the predictive variance is computed as $v_*^{(k)} = \mathbf{k}_{\mathbf{x}_*}^{+^T} \Sigma_q^{(k)^+} \mathbf{k}_{\mathbf{x}_*}^+$, where $\mathbf{k}_{\mathbf{x}_*}^+ = (\mathbf{k}_{\mathbf{x}_*}^T, k(\mathbf{x}_*, \hat{\mathbf{x}}_1), ..., k(\mathbf{x}_*, \hat{\mathbf{x}}_S))^T$, with a reduced computational cost of $O((M+S)^2)$.

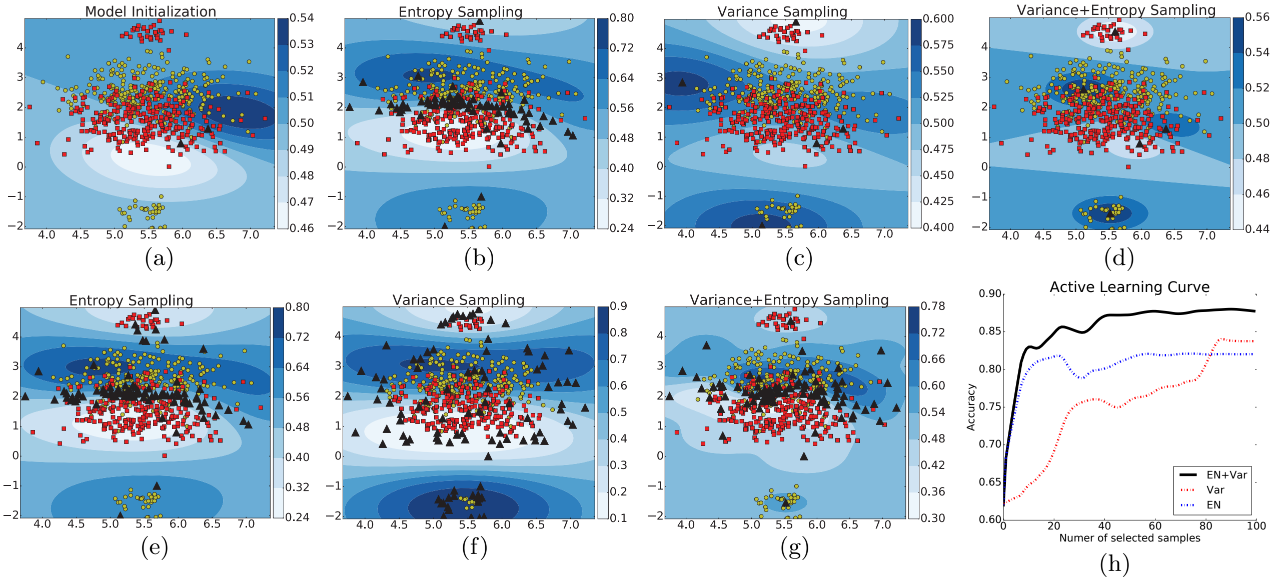
4 Experiments

We present our experimental results over both synthetic and real-world data to demonstrate: (1) the effective active sampling behavior of MM-SGP through the dynamic balancing between predicted variance and entropy, (2) overall better AL performance than other competitive multi-class AL models, and (3) how sparsity ensures good AL performance while significantly accelerating model retraining.

4.1 Synthetic Data Experiments

We generate a 2D synthetic dataset to help demonstrate important model properties and sampling behaviour of the proposed MM-SGP. There are two clusters of the positive and negative samples highly overlapped in the center region. Each main cluster is close to a small cluster of samples from the opposite class. We initialize the model by simulating a typical starting point of AL, where the initial training data (shown as the black triangles in Figure 2(a)) is very likely to come from most densely distributed area of the data space. As a result, the initial decision boundary can be highly incorrect, as clearly demonstrated by Figure 2(a). Next, we compare the sampling behavior of three sampling strategies at different critical phases of AL, including entropy based, variance based, and MM-SGP, which combines and dynamically balances between variance and entropy.

First, limited labeled data usually lead to insufficient initialization of an AL model. Therefore, effective exploration is critical in the early stage of AL and the model performance cannot be significantly improved until it sufficiently explores the entire data space and discovers areas that may dramatically change the decision boundary. We observe that the entropy-based sampling exhibits a random exploration behaviour because the entropy is high on both the current decision boundary and areas that away from the training data. Figure 2(b) shows that by the time the model discovers both small clusters, 17 and 78 additional data samples are labeled. These also correspond to the two locations on the entropy AL curve in Figure 2(h) that trigger a fast improvement of the model accuracy. In most of other times, the model myopically samples many data points near an incorrect decision boundary, which



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2: Decision boundaries and sampled data distribution resulted from predicted entropy, variance, and their combination: (a) model initialization; (b)–(d) towards the end of the exploration stage, which is early in AL; (e)–(g) toward the end of the entire AL process; h) overall AL curves for three sampling methods.

Table 1: Summary of Datasets

Dataset	Dermatology I	Dermatology II	Yeast	Penstroke	Auto-drive	Reuters
Domain	Medical	Medical	Biology	Image	Driving	News
Sample	800	868	1484	1144	58509	10788
Feature	1391	1554	8	500	48	5227
Class	50	30	10	26	11	75

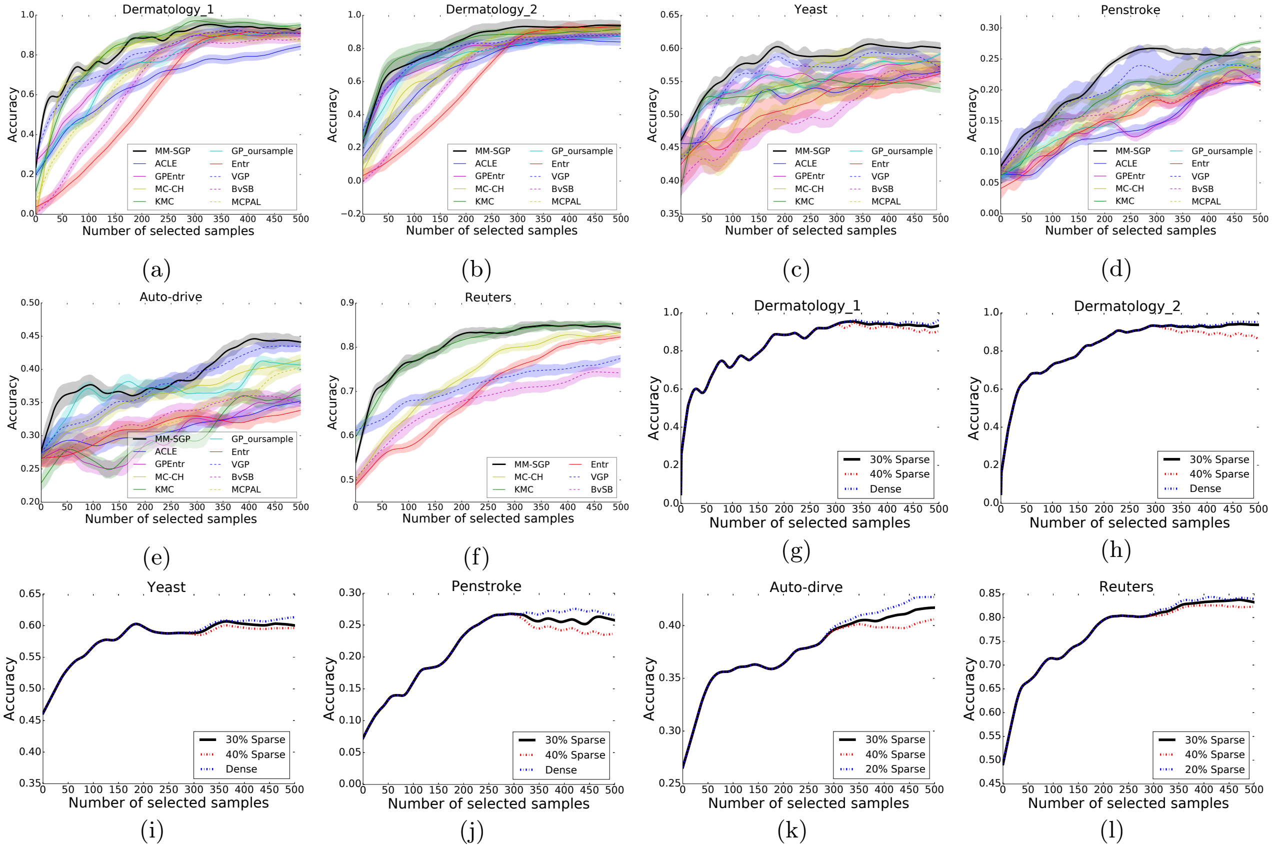
does not contribute a lot to the model improvement. In contrast, variance-based sampling is able to more effectively explore the data space. Figure 2(c) shows that the model only takes three iterations to discover the two small clusters and learns a roughly correct decision boundary. However, Figure 2(f) indicates that the model keeps sampling on the edge of the feature data and failed to fine-tune the decision boundary in the overlapping areas thus resulting in a lower convergence accuracy as shown by the Var curve in Figure 2(h).

The proposed MM-SGP AL model dynamically balances the predicted variance and entropy to effectively explore the data space in early AL (Figure 2 (d)) while performing sufficient fine-tuning in later AL (Figure 2 (g)) to achieve much more efficient convergence to a higher model accuracy as shown by the EN+Var curve in Figure 2 (h).

4.2 Experiments on Real-World Data

We choose six representative real-world datasets from different domains, where the number of classes vary from 10 to 75. The details of those datasets are shown in Table 1. To simulate the real-world AL tasks with very limited initial labeled data, we start AL with one labeled sample from each class. Due to the small initial training size, we transition to the sparse GP when the number of samples reaches 300. For the comparison experiments, we set the sparse level at 30% ($M = 0.7N$). We initialize η as 10 and dynamically decrease it along with AL so that it reaches zero in the end. We will study the impact of η in more details later. Coefficient c in (1) is cross-validated and set to 0.01. We adopt the linear kernel for Reuters and RBF kernel for the other datasets.

Performance comparison. We compare the proposed MM-SGP model with several competitive AL models from three different categories, including (1) SVM-based: Best-verus-Second Best (BvSB) [22], convex-hull based unified sampling (MC-CH) [26], Entropy-based sampling (Entr); (2) GP-based: a standard GP that uses entropy for sampling (GPEntr), a standard GP that uses the proposed sampling function (GP_oursample), our model with sparsity and maximum margin constraints removed (VGP), kernel committee machine (KMC) model [17], which combines a parametric GP (i.e., RVM) with an SVM for AL; and (3) specially designed classification models for AL: multi-class probabilistic active learning (McPAL) [25], AL with cost embedding (ACLE) [31]. The

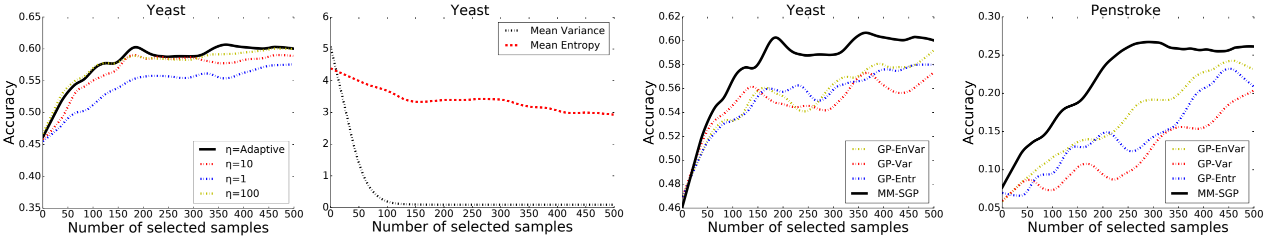


(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 3: Comparison with other AL models (a)–(f); Effectiveness of model sparsity (g)–(l)



(a) (b) (c) (d)

Figure 4: Impact of the balancing parameter η (a)-(b) and MM constraints (c)-(d);

first three models use SVMs as the active learner, which also leverage the max-margin constraints for learning as our model does. We do not provide the result for McPAL, ACLE, and GP on Reuters because their very high computational cost in the later phase of AL.

Figure 3 (a)-(e) show the AL curves (with variance) for each dataset. It can be seen that MM-SGP has a clear advantage on most of the datasets in the beginning of AL. We also notice that AL models solely rely on entropy for sampling (e.g., GPEntr and Entr) exhibit no such behavior. This further justifies the effectiveness of early exploration guided by an accurate predictive variance. The performance advantage over the KMC model also confirms that MM-SGP can more accurately predict the variance information for more effective data sampling than using a degenerate GP (i.e., RVM).

After the fast improvement at the early stage of AL, the performance of MM-SGP starts to grow steadily and maintain its leading position towards the end of AL. In Figure 4, we provide further insight on where the predictive entropy starts to take a lead in sampling and the model starts to exploit its decision boundary determined by the early sampled data through effective exploration. In addition to the effective sampling behavior, comparison with a standard GP using our sampling function and VGP helps to demonstrate the contribution of the MM constraints that lead to a higher prediction performance even with a sparse GP model.

Impact of sparsity. Figure 3 (g)-(i) demonstrate the AL behavior at different levels of sparsity. For

Table 2: Sampling Time Comparison

Trainsize	300				400				500			
Sparsity	75%	50%	30%	dense	75%	50%	30%	dense	75%	50%	30%	dense
Yeast	1.1	6	11.7	29.9	3.4	14.3	28.9	78.9	8.7	41	108.1	282.3
Dermatology I	7.1	23.6	41.7	251.7	31.1	76.3	216.5	844.5	46.6	202.1	545.9	886.6
Dermatology II	5.6	17.7	30.2	77.6	13.1	52.3	120.6	416.8	26.8	98.1	287.5	1019.3
Penstroke	7.2	25.7	46.9	177.3	22.4	42.9	103.6	315.2	21.5	83	306.2	893.3
Auto-drive	4.7	12.1	20.8	86	7.1	25.7	60	194.1	13	48.6	117	286
Reuters	55	110.5	164.2	486.7	74	159.9	481.4	1003.2	103	263	672.1	1295.3

Reuters and Auto-drive, the dense model cannot complete within a reasonable time for AL purpose, so we report the learning curve at 20% sparsity. As can be seen, for most cases, 30% sparsity achieves almost the same AL performance as the dense model. Further increasing the sparsity (e.g., 40%) may hurt model performance in a few cases, including Penstroke and Dermatology II. For Auto-drive, a 40% sparse level still performs very well but is slightly less effective than a denser model.

Sampling time comparison. Besides being able to maintain a very competitive AL performance, the MM-SGP can be trained much more efficiently to achieve a 3-5X speed-up than the dense version. Table 2 reports the sampling time (in seconds) of MM-SGP at different level of sparsity. The result shows that as the model becomes more and more sparse, the execution speed of MM-SGP is up to 30X faster than the dense GP. However, maintaining the highest sparsity will hurt the classification performance, especially when the train size is small. In most of our experiments, we adopt the sparsity level of 30% to achieve a good balance between execution efficiency and model accuracy.

Impact of key model parameters. Since the AL performance is quite stable with greater than 5 augmented samples, we fix S to 10 for all the datasets. Figure 4 (a)-(b) use Yeast as an example to demonstrate how η affects data sampling (results for other datasets are in Appendix D). In particular, we compare our sampling method that adaptively decays η along with AL to balance entropy and variance with some fixed η values. The result shows, by setting $\eta = 10$, we achieve an AL curve most close to the adaptive η. Other η values achieve slightly worse but quite comparable performance. The underlying reason is that the predictive variance of the entire candidate pool reduces automatically as MM-SGP continues to explore the data space effectively. This can be seen from Figure 4 (b) where the mean predictive variance and entropy over the candidate pool are re-scaled and plotted against each other. The predictive variance quickly drops to almost zero as the model explores the data space at the early stage of AL. In contrast, the predictive entropy steady decreases as AL moves forward but never approaches zero due to the overlapping of different classes.

Impact of MM constraints. To further demonstrate the effectiveness of integrating the MM constraints, we compare the proposed model with a standard GP. Since a GP can output both the predictive entropy and variance using the approaches as we described, we pair a GP with three sampling methods that rely on predictive entropy, variance, and their combination. Figure 4 (c)-(d) show the comparison results using Yeast and Penstroke as examples (results from other datasets are provided in Appendix D). MM-SGP achieves a much better performance both along the AL process and upon convergence, which clearly demonstrates the effectiveness of the MM constraints.

5 Conclusion

We propose a Maximum-Margin Sparse Gaussian Process (MM-SGP) for AL in multi-class classification. By leveraging the parametric view of a GP, we seamlessly integrate a nonparametric GP into the maximum entropy discrimination framework, leading to the improved discriminate power of MM-SGP with a small "effective size". Augmentation using the unlabeled data systematically fixes the underestimated predictive variance, which is critical for data sampling. Furthermore, the sparse nature of MM-SGP ensures that the convex dual problem for posterior inference has a nice low-rank structure, which can be efficiently solved to support real-time expert-machine collaborative AL. A novel active sampling function is further developed to choose the most effective data samples for model training by dynamically balancing the predicted variance and entropy. Comparison with competitive multi-class AL models clearly demonstrate its effectiveness and performance advantage in real-world AL tasks from diverse domains.

Acknowledgements

This research was supported in part by an NSF IIS award IIS-1814450 and an ONR award N00014-18-1-2875. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agency.

References

[1] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[2] Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 917–926. ACM, 2009.

[3] Oisin Mac Aodha, Neill DF Campbell, Jan Kautz, and Gabriel J Brostow. Hierarchical subquery evaluation for active learning on a graph. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 564–571, 2014.

[4] Meng Wang and Xian-Sheng Hua. Active learning in multimedia annotation and retrieval: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(2):10, 2011.

[5] Tommi Jaakkola, Marina Meila, and Tony Jebara. Maximum entropy discrimination. In *Advances in neural information processing systems*, pages 470–476, 2000.

[6] Jun Zhu and Eric P Xing. Maximum entropy discrimination markov networks. *Journal of Machine Learning Research*, 10(Nov):2531–2569, 2009.

[7] J Quinonero Candela and Lars Kai Hansen. Learning with uncertainty-gaussian processes and relevance vector machines. *Technical University of Denmark, Copenhagen*, 2004.

[8] Sanjoy Dasgupta. Two faces of active learning. *Theoretical computer science*, 412(19):1767–1781, 2011.

[9] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959, 2005.

[10] David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

[11] David A Cohn. Neural network exploration using optimal experiment design. In *Advances in neural information processing systems*, pages 679–686, 1994.

[12] Sambu Seo, Marko Wallat, Thore Graepel, and Klaus Obermayer. Gaussian process regression: Active data selection and test point rejection. In *Mustererkennung 2000*, pages 27–34. Springer, 2000.

[13] Andreas Krause and Carlos Guestrin. Nonmyopic active learning of gaussian processes: an exploration-exploitation approach. In *Proceedings of the 24th international conference on Machine learning*, pages 449–456, 2007.

[14] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Active learning with gaussian processes for object categorization. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[15] Ashish Kapoor, Eric Horvitz, and Sumit Basu. Selective supervision: Guiding supervised learning with decision-theoretic active learning. In *IJCAI*, volume 7, pages 877–882, 2007.

[16] Catarina Silva and Bernardete Ribeiro. Combining active learning and relevance vector machines for text classification. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pages 130–135. IEEE, 2007.

[17] Weishi Shi and Qi Yu. Integrating bayesian and discriminative sparse kernel machines for multi-class active learning. In *Thirty-third Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[18] Carl Edward Rasmussen and Joaquin Quinonero-Candela. Healing the relevance vector machine through augmentation. In *Proceedings of the 22nd international conference on Machine learning*, pages 689–696. ACM, 2005.

[19] Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *ICML*, pages 839–846, 2000.

[20] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

[21] Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 5, pages 746–751, 2005.

[22] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379. IEEE, 2009.

[23] Feng Jing, Mingjing Li, HongJiang Zhang, and Bo Zhang. Entropy-based active learning with support vector machines for content-based image

retrieval. In *IEEE International Conference on Multimedia & Expo (ICME)*, pages 85–88, 2004.

[24] Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.

[25] Daniel Kottke, Georg Krempl, Dominik Lang, Johannes Teschner, and Myra Spiliopoulou. Multi-class probabilistic active learning. In *European Conference on Artificial Intelligence (ECAI)*, pages 586–594, 2016.

[26] Weishi Shi and Qi Yu. An efficient many-class active learning framework for knowledge-rich domains. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 1230–1235. IEEE, 2018.

[27] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[28] Shai Fine and Katya Scheinberg. Efficient svm training using low-rank kernel representations. *Journal of Machine Learning Research*, 2(Dec):243–264, 2001.

[29] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[30] Grace Wahba, Xiwu Lin, Fangyu Gao, Dong Xiang, Ronald Klein, and Barbara Klein. The bias-variance tradeoff and the randomized gacv. In *Advances in Neural Information Processing Systems*, pages 620–626, 1999.

[31] Kuan-Hao Huang and Hsuan-Tien Lin. A novel uncertainty sampling algorithm for cost-sensitive multiclass active learning. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 925–930. IEEE, 2016.

[32] Yao-Yuan Yang, Shao-Chuan Lee, Yu-An Chung, Tung-En Wu, Si-An Chen, and Hsuan-Tien Lin. libact: Pool-based active learning in python. Technical report, National Taiwan University, October 2017.