## Reply to Reviewer 1

We thank the reviewer for the constructive suggestion and recommendation for acceptance. We appreciate your positive evaluation on the novelty of the proposed unified AL model along with the sampling function and the convincing experimental results. We also agree with the reviewer about the importance of generalization guarantees and label complexity bounds, which will be an exciting direction that we plan to pursue in our future work. We would like to point out that the proposed model introduces a much more complex (regularized) loss function in order to integrate two different types of sparse kernel machines for more effective sampling and uses a specially designed sampling function for multi-class problems. Addressing these challenges in the theoretical analysis could help further extend the state of the art on the theory of active learning.

## Reply to Reviewer 2

We thank the reviewer for the thoughtful feedback and recommendation for acceptance. We appreciate your positive evaluation on the novelty of the optimization algorithm for the new learning objective and the AL policy.

*Q1: RVM is a probabilistic, not a generative model.* We agree that RVM does not model $p(x|y)$ and our initial intent is to leverage RVM's capability to model the conditional distribution of the response $y$ given the input $x$. We will make this clear as suggested by the reviewer.

*Q2: The learning curves started from different accuracy and not all active learning methods use KMC.* The different starting accuracy is caused by different learning models. In fact, for many AL methods, the sampling rules are designed for specific learning algorithms. In our proposed approach, the sampling rule given in eq. (13) is developed along with the KMC model as it uses model-specific information for sampling. Due to this coupling, the selected data sample can help improve the given model to the largest extent. The same rationale also applies to several other competitive models in our experiments. For example, MC-CH is built upon SVM as it uses the convex hull of support vectors for sampling. Similarly, McPAL requires its own learning model and BvSB is typically used with an SVM model.

*Q3: Should report average or median results in experiments.* The reported test accuracy is averaged over three runs. We will make this clear in the revised paper.

*Q4: How large is S in the experiments on the real datasets.* S is set to 40, which will be made clear.

*Q5: Compare RVM, SVM, and KMC in passive learning setting.* Following the reviewer's suggestion, we have compared these models in passive learning using the 6 real-world datasets. The general trend is that with limited training data, RVM and KMC perform better than SVM as SVM may be easily trapped to a local optimal decision boundary. With sufficient training data, SVM and KMC achieve comparable model performance and both outperform RVM. However, SVM requires a large number of support vectors to fine-tune the decision boundary while KMC uses much less KMC vectors. In summary, in passive learning, KMC can automatically adapt to the size of the training data and provide robust and competitive classification performance in all cases, which mainly benefits from the unified objective function.

## Reply to Reviewer 3

We thank the reviewer for the constructive feedback. We appreciate your positive evaluation on the novelty of the objective function/lower bounding of the optimization problem and thorough experimental results.

*Q1: Theorem 2 should be made much more formal.* We will provide a more formal statement and proof of the Theorem (and also fix the label) as suggested by the reviewer. In particular, the Theorem can be more formally specified as " Using an ARD prior, the covariance matrix $S_q$ of variational distribution $q(\mathbf{w})$ has a sparse structure. In particular, for $|\alpha_i| \to \infty$ and $|\alpha_j| \to \infty$, $S_q(i,j) \to 0$ as $S_q(i,j) \propto 1/|\alpha_j|$ (similarly $S_q(j,i) \to 0$ as $S_q(j,i) \propto 1/|\alpha_i|$)". We will provide a reference to Faul, A. C., & Tipping, M. E. (2002), which proved that some $\alpha$'s approach $\infty$ to ensure the sparsity of RVM (also verified in our experiments). We will also provide more details for the proof to make it clear that $S_q(i,j) \propto 1/|\alpha_j|$ for $|\alpha_i|, |\alpha_j| \to \infty$. A key step added to the current proof is to apply the Woodbury identity to the term $(\Lambda^{-1} + \Phi A^{-1}\Phi)^{-1}$ in eq.(17) and by using the fact that $|\alpha_i| \to \infty$, we can show $(\Lambda^{-1} + \Phi A^{-1}\Phi)^{-1} \approx A$. Using this fact and eq. (18), we can show $S_q(i,j) \propto 1/|\alpha_j|$ and hence $S_q(i,j) \to 0$ for $|\alpha_j| \to \infty$.

*Q2: Derivation of (1) is difficult to follow.* The key insight of objective function (1) is to combine a likelihood term that well captures the data distribution with a large margin constraint to simultaneously ensure good discriminative power of the model. The regularizer is added to help ensure model sparsity.

*Q3: It is not entirely clear that simulations in Figures 2 and 3 show the authors' claim.* The main purpose of these figures is to show that KMC sufficiently explores critical areas of the data distribution while giving adequate attention to the decision boundaries by using limited KMC vectors. While KMC uses slightly more vectors than RVM, it is much sparser than SVM. The middle chart of Figure 3 shows excessive support vectors are assigned close to the decision boundary (including some low density areas) while KMC only assigns a few vectors there as shown in the right chart.