The Multiple Instance Learning Gaussian Process Probit Model

Fulton Wang
Sandia National Laboratories
fulwang@sandia.gov

Ali Pinar
Sandia National Laboratories
apinar@sandia.gov

Abstract

In the Multiple Instance Learning (MIL) scenario, the training data consists of instances grouped into bags. Bag labels specify whether each bag contains at least one positive instance, but instance labels are not observed. Recently, Haußmann et al [10] tackled the MIL instance label prediction task by introducing the Multiple Instance Learning Gaussian Process Logistic (MIL-GP-Logistic) model, an adaptation of the Gaussian Process *Logistic* Classification model that inherits its uncertainty quantification and flexibility. Notably, they give a fast mean-field variational inference procedure. However, due to their use of the logit link, they do not maximize the variational inference ELBO objective directly, but rather a *lower bound* on it. This approximation, as we show, hurts predictive performance. In this work, we propose the Multiple Instance Learning Gaussian Process Probit (MIL-GP-Probit) model, an adaptation of the Gaussian Process *Probit* Classification model to solve the MIL instance label prediction problem. Leveraging the analytical tractability of the probit link, we give a variational inference procedure based on variable augmentation that maximizes the ELBO objective *directly*. Applying it, we show MIL-GP-Probit is more calibrated than MIL-GP-Logistic on all 20 datasets of the benchmark 20 Newsgroups dataset collection, and achieves higher AUC than MIL-GP-Logistic on an additional 51 out of 59 datasets. Finally, we show how the probit formulation enables principled bag label predictions and a Gibbs sampling scheme. This is the first exact inference scheme for *any* Bayesian model for the MIL scenario.

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

	MIL-GP-Probit (ours)	MIL-GP-Logistic ([10])
Predictive likelihood wins on 20 Newsgroup dataset collection	**20**	0
AUC wins on 59 datasets from [18]	**51**	8

Table 1: Our MIL-GP-Probit model beats the state-of-the-art MIL-GP-Logistic model of [10] on the majority of datasets in 2 dataset collections in terms of instance label predictive likelihood and AUC, due to our more accurate variational inference method.

1 Introduction

In the Multiple Instance Learning (MIL) scenario [4], the training data consists of instances grouped into bags. Each instance has a binary instance label, but it is unobserved. Instead, each bag is labelled with a bag label according to the MIL labeling assumption: a bag label is positive if and only if the bag contains at least one positive instance. There are two possible tasks in the MIL scenario: the bag label prediction problem, and the instance label prediction problem. The latter task is arguably harder, and more common. In computer vision [22], one wants to classify whether a patch is a dog, given training data that only specifies whether each image contains at least one dog patch. In activity recognition [8], one wants to classify whether a person was exercising during a short time window, given training data that only specifies whether a person was exercising at any time within a long time window.

In this work, we tackle the MIL *instance* label prediction problem. Prior work on this problem has included maximum margin methods [2, 26] and probabilistic methods [19, 24, 5], which explicitly model the probability instance labels are positive. A key aspect of the MIL scenario is that there is lots of *model uncertainty*, because there will be many models plausible given the ambiguous training data. Consequently, a series of Bayesian methods for the MIL scenario was developed to explicitly account for model uncertainty, including those based on Bayesian linear regression

[20], Dirichlet Process Mixture Models [13], and Gaussian Processes [14]. However, these methods all have drawbacks. [20, 14] use Laplace's approximation for inference, which is slow and inaccurate; [13] uses a generative model, making it suitable only for low-dimensional data, and [14] does not model instance labels, and thus cannot solve the instance label prediction problem.

Recently, [10] developed MIL-GP-Logistic, an adaptation of the Gaussian Process Logistic Classification model to solve the MIL instance label prediction problem that does not suffer from any of these drawbacks. Their model is discriminative, offering the nonparametric flexibility of the Gaussian Process model, and explicitly models instance labels so that it can actually solve the instance label prediction problem. Furthermore, they give a mean-field variational inference procedure that has closed-form updates. Applying the inference procedure, their model achieves state of the art performance on several benchmark datasets. Yet, MIL-GP-Logistic is not without flaws: due to their choice of the logit link[1], they do not maximize the standard ELBO variational inference objective directly, but rather a lower bound on it. As our experiments show, this approximation hurts predictive performance.

To address this drawback, we propose the MIL-GP-Probit model, an adaptation of the Gaussian Process *Probit* Classification model to solve the MIL instance label prediction problem. Leveraging the analytical tractability of the probit link, we provide a variational inference procedure based on variable augmentation that maximizes the ELBO *directly*, instead of a lower bound on it. Applying it, we show MIL-GP-Probit has better predictive performance than MIL-GP-Logistic on many MIL datasets (see Table 1).

Furthermore, we develop a Gibbs sampling inference scheme for the MIL-GP-Probit model. To the best of our knowledge, this is the first (asymptotically) exact inference scheme for *any* Bayesian model for the MIL scenario. Comparing the approximate posterior given by our variational inference scheme to the exact posterior given by our Gibbs sampling scheme, we find that the former sacrifices little, if any, predictive performance. Finally, although our focus is on predicting instance labels, we also provide a principled way to make bag label predictions that accounts for the *dependence* between instance labels asserted by the GP model. The model of [10] lacks this capability, which would be crucial for active learning, where the bag label that the model is most uncertain about might be requested.

Our contributions are as follows: We 1) develop MIL-GP-Probit, a Gaussian Process model for the MIL instance label prediction problem which differs from the state-of-the-art MIL-GP-Logistic model of [10] by a crucial design choice: the use of a probit link instead of logit link; leverage the analytical tractability of the probit link to 2) develop a mean-field variational inference scheme for MIL-GP-Probit that maximizes the ELBO *directly*, instead of a lower bound as did [10] and 3) develop a Gibbs sampling inference scheme for MIL-GP-Probit - the first exact inference scheme for *any* Bayesian model for the MIL scenario; 4) show that MIL-GP-Probit with variational inference is significantly more calibrated than MIL-GP-Logistic on all 20 datasets of the benchmark 20 Newsgroups dataset collection, achieves higher AUC than MIL-GP-Logistic on an additional 51 out of 59 datasets, and gives predictive performance comparable to Gibbs sampling.

2 Background

Multiple Instance Learning: In the Multiple Instance Learning (MIL) scenario, the training data consists of N instances partitioned into a set of bags $\mathcal{B}$, i.e. $b = \{b_1, \ldots, b_{|b|}\} \subseteq [N]$ for $b \in \mathcal{B}$, where $|b|$ is the size of bag b. Bag b contains $|b|$ instances $X_b \coloneqq \{x_i\}_{i \in b}$, with $x_i \in \mathbb{R}^D$ where D is the number of features, and associated binary instance labels $H_b \coloneqq \{h_i\}_{i \in b}$, $h_i \in \{0, 1\}$ for $i \in [N]$. The instance labels $\{H_b\}$ are not observed. Instead, one observes bag labels $\{Y_b\}$. The MIL labeling assumption is that $Y_b = 1$ if and only if any instance label in bag b is positive. That is, $Y_b = \max_{i \in b} h_i$. Thus, the training data comprises labelled bags $\{X_b\}_{b \in \mathcal{B}}, \{Y_b\}_{b \in \mathcal{B}}$. Probabilistic MIL methods create a model $P(\{Y_b\}, \{H_b\} | \{X_b\})$ that factors as $P(\{H_b\}|\{X_b\})\Pi_{b \in \mathcal{B}} P^{\text{MIL}}(Y_b|H_b)$. They vary by how they model $P(\{H_b\}|\{X_b\})$. However, the MIL labeling assumption dictates that

$$P^{\text{MIL}}(Y_b|H_b) = 1[Y_b = \max_{i \in b} h_i]. \tag{1}$$

Gaussian Process Classification:. Our Multiple Instance Learning Gaussian Process model is an adaptation of a Gaussian Process (GP) classification model, which is itself an adaptation of a Gaussian Process model, which we describe first. Given instance features $\mathbf{x} = (x_1, \ldots, x_N)$, instance scalars $\mathbf{f} = (f_1, \ldots, f_N)$, where $f_n \in \mathbb{R}$, $x_n \in \mathbb{R}^D$ for $n \in [N]$ and D is the number of features, and kernel function $K(\cdot, \cdot; \theta)$ parametrized by kernel hyperparameters θ, a Gaussian Process (GP) model lets $\mathbf{f}; \mathbf{x}, \theta \sim \mathcal{N}(\mathbf{0}_N, K_{\mathbf{xx};\theta})$, where $\mathbf{0}_N$ is the length N vector of zeros. Throughout, given $\mathbf{x}' = (x_1, \ldots, x_{N'})$, $\mathbf{x}'' = (x_1, \ldots, x_{N''})$ and kernel function $K(\cdot, \cdot; \theta)$, we overload $K_{\mathbf{x}'\mathbf{x}'';\theta}$ to denote the $N' \times N''$ gram matrix whose (i, j)-th entry is $K(x'_i, x''_j; \theta)$.

1. Logistic classification models use a logit link (equivalently, logistic inverse link), whereas probit classification models use a probit link, leading to unfortunate inconsistency in terminology.

When $\mathbf{f}$ is observed, for test instances with features $\mathbf{x}^*$ and instance scalars $\mathbf{f}^*$, the posterior predictive distribution $\mathbf{f}^*|\mathbf{x}^*, \mathbf{f}, \mathbf{x}; \theta \sim \mathcal{N}(\mathbf{f}^* K_{\mathbf{x}^*\mathbf{x};\theta} K^{-1}_{\mathbf{xx};\theta} \mathbf{f}, K_{\mathbf{x}^*\mathbf{x}^*;\theta} - K_{\mathbf{x}^*\mathbf{x};\theta} K^{-1}_{\mathbf{xx};\theta} K_{\mathbf{xx}^*;\theta})$. Due to the $O(N^3)$ cost of inverting a $N \times N$ matrix, the Fully Independent Training Conditional (FITC) approximation [21] introduces R inducing points $\mathbf{z} \coloneqq (z_1, \ldots, z_R)$ and inducing scalars $\mathbf{u} \coloneqq (u_1, \ldots, u_R)$, where $z_r \in \mathbb{R}^D$, $u_r \in \mathbb{R}$ for $r \in [R]$, and lets

$$\mathbf{u}; \mathbf{z}, \theta \sim \mathcal{N}(\mathbf{0}_R, K_{\mathbf{zz};\theta}), \tag{2}$$

$$\mathbf{f}|\mathbf{u}; \mathbf{x}, \mathbf{z}, \theta \sim \mathcal{N}(K_{\mathbf{xz};\theta} K^{-1}_{\mathbf{zz};\theta} \mathbf{u}, \operatorname{diag}(\mathcal{K})), \tag{3}$$

where $\mathcal{K} \coloneqq K_{\mathbf{xx};\theta} - K_{\mathbf{xz};\theta} K^{-1}_{\mathbf{zz};\theta} K_{\mathbf{zx};\theta}$. and $\operatorname{diag}(\mathcal{K})$ is the diagonal matrix whose diagonal equals that of $\mathcal{K}$. As the covariance of $P(\mathbf{f}|\mathbf{u}; \mathbf{x}, \mathbf{z}, \theta)$ is $\operatorname{diag}(\mathcal{K})$ instead of $\mathcal{K}$, inference cost is reduced to $O(R^2 N)$. Throughout, we omit notational dependence on $\mathbf{z}, \theta$ if appropriate.

The Gaussian Process *Logistic* Classification model (with FITC approximation) extends the model of Equations 2 and 3 by additionally modeling binary labels $\mathbf{h} \coloneqq (h_1, \ldots, h_N)$, letting

$$h_i \sim \textsc{Bernoulli}(\textsc{logit}^{-1}(f_i)) \text{ for } i \in [N], \tag{4}$$

where $\textsc{logit}^{-1}(x) = \frac{1}{1+\exp(-x)}$ is the logistic function, so that a logit link (equivalently, logistic inverse link) is used.

Similarly, the Gaussian Process *Probit* Classification model (with FITC approximation) extends the model of Equations 2 and 3 by additionally modeling binary labels $\mathbf{h} \coloneqq (h_1, \ldots, h_N)$, but instead letting

$$h_i \sim \textsc{Bernoulli}(\textsc{probit}^{-1}(f_i)) \text{ for } i \in [N],$$

where $\textsc{probit}^{-1}(z) = \int 1[x < z] \mathcal{N}(x; 0, 1) dx$ is the cumulative distribution function of the standard normal, so that a probit link is used.

3 Multiple Instance Learning Gaussian Process Probit Model

3.1 Model Formulation

Our Multiple Instance Learning Gaussian Process Probit model (MIL-GP-Probit) models the MIL setting using the probabilistic approach described in Section 2, where the model for instance labels $P(\{H_b\}|\{X_b\})$ uses a Gaussian Process *Probit* Classification model (with FITC approximation). Given R inducing points $\mathbf{z} \coloneqq (z_1, \ldots, z_R)$ with $z_r \in \mathbb{R}^D$ and kernel function $K(\cdot, \cdot; \theta)$ depending on hyperparameters θ, the MIL-GP-Probit model for the MIL setting models

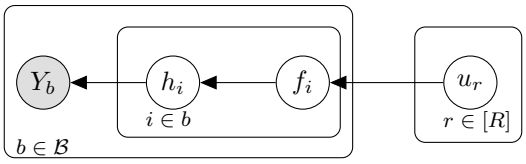

Figure 1: In the MIL-GP-Probit model, instance labels $\{h_i\}$ are modeled using a Gaussian Process Probit Classification model (with FITC approximation), and related to bag labels $\{Y_b\}$ using the MIL assumption.

$P(\mathbf{u}, \mathbf{f}, \{H_b\}, \{Y_b\}; \{X_b\}, \mathbf{z}, \theta)$ as:

$$\mathbf{u}; \mathbf{z}, \theta \sim \mathcal{N}(\mathbf{0}_R, K_{\mathbf{zz};\theta}), \tag{5}$$

$$\mathbf{f}|\mathbf{u}; \mathbf{x}, \mathbf{z}, \theta \sim \mathcal{N}(K_{\mathbf{xz};\theta} K^{-1}_{\mathbf{zz};\theta} \mathbf{u}, \operatorname{diag}(\mathcal{K})), \tag{6}$$

$$h_i \sim \textsc{Bernoulli}(\textsc{probit}^{-1}(f_i)) \text{ for } i \in [N], \tag{7}$$

$$Y_b|H_b \sim 1[Y_b = \max_{i \in b} h_i] \text{ for } b \in \mathcal{B}, \tag{8}$$

where $\mathbf{x}$ denotes $\{X_b\}$, $\mathbf{f} \in \mathbb{R}^N$ denotes $\{F_b\}$ where $F_b \coloneqq \{f_i\}_{i \in b}$ and the f_i are instance scalars, $\mathbf{u} \coloneqq (u_1, \ldots, u_R)$ where $u_r \in \mathbb{R}$ are inducing scalars, and $\mathcal{K} \coloneqq K_{\mathbf{xx};\theta} - K_{\mathbf{xz};\theta} K^{-1}_{\mathbf{zz};\theta} K_{\mathbf{zx};\theta}$. Equations 5 - 7 specify the Gaussian Process Probit Classification model (with FITC approximation) for all instance labels $\{H_b\}$ given all instance features $\mathbf{x} \coloneqq \{X_b\}$, and Equation 8 specifies the MIL labeling assumption that holds between a given bag label Y_b and the instance labels of the bag H_b. Figure 1 depicts the MIL-GP-Probit model.

3.2 Model Inference

Given bag labels $\{Y_b\}$, the inference task is to compute $P(\mathbf{u}, \mathbf{f}|\{Y_b\}; \mathbf{x}, \mathbf{z}, \theta)$. A variable augmentation approach enables efficient variational inference of this posterior.

Variable Augmentation: We modify the MIL-GP-Probit model to obtain the *augmented* MIL-GP-Probit model via 2 transforms.

In the 1st transform, we introduce augmenting variable $\mathbf{m} \coloneqq (m_1, \ldots, m_N)$, with $m_i \in \mathbb{R}$, and we will define $M_b \coloneqq \{m_i\}_{i \in b}$. Then for $i \in [N]$ we let

$$m_i \sim \mathcal{N}(f_i, 1) \tag{9}$$

$$h_i \sim 1[m_i > 0]. \tag{10}$$

The marginal distribution of the original variables is identical between the original MIL-GP model and the transformed model, as under the latter, for $i \in [N]$,

$$\begin{aligned} P(h_i = 1|f_i) &= \int P(h_i = 1|m_i) P(m_i|f_i) dm_i \\ &= \int 1[m_i > 0] \mathcal{N}(m_i; f_i, 1) dm_i \\ &= \int 1[f_i - x > 0] \mathcal{N}(f_i - x; f_i, 1) dx \\ &= \int 1[x < f_i] \mathcal{N}(x; 0, 1) dx \\ &= \textsc{probit}^{-1}(f_i), \end{aligned}$$

so it remains that $h_i|f_i \sim \text{BERNOULLI}(\text{PROBIT}^{-1}(f_i))$.

In the 2nd transform, we marginalize out $\{H_b\}$ so that

$$\begin{aligned} P(Y_b = 0|M_b) &= \sum_{H_b \in \{0,1\}^{|b|}} P(Y_b = 0|H_b)P(H_b|M_b) \\ &= \sum_{H_b \in \{0,1\}^{|b|}} 1[0 = \max_{i \in b} h_i] \Pi_{i \in b} 1[h_i > m_i] \\ &= \Pi_{i \in b} 1[0 > m_i], \text{ and} \quad (11) \\ P(Y_b = 1|M_b) &= 1 - \Pi_{i \in b} 1[0 > m_i]. \quad (12) \end{aligned}$$

The result of these 2 transforms is the *augmented* MIL-GP-Probit model $P(\mathbf{u}, \mathbf{f}, \mathbf{m}, \{Y_b\}; \mathbf{x})$ given by Equations 5, 6, 9, 11, and 12.

This variable augmentation approach enabled Gibbs sampling for binary probit models [1] and variational inference for multi-class Gaussian Process probit models [7]. As we now show, it also enables variational inference for our MIL-GP-Probit model.

Variational Inference: We use mean-field variational inference (VI) to efficiently approximate the posterior of the augmented MIL-GP-Probit model, $P(\mathbf{u}, \mathbf{f}, \mathbf{m}|\{Y_b\}; \mathbf{x})$. We approximate this intractable posterior with a variational distribution $Q(\mathbf{u}, \mathbf{f}, \mathbf{m}) \in \mathcal{Q}$, where $\mathcal{Q}$ are distributions factorizing as:

$$Q(\mathbf{u}, \mathbf{f}, \mathbf{m}) = Q(\mathbf{u})Q(\mathbf{m})P(\mathbf{f}|\mathbf{u}; \mathbf{x}) \quad (13)$$

VI seeks $\text{argmin}_{Q \in \mathcal{Q}} KL(Q||P(\mathbf{u}, \mathbf{f}, \mathbf{m}|\{Y_b\}; \mathbf{x}))$. This is $\text{argmax}_{Q \in \mathcal{Q}} \text{ELBO}(Q)$, with

$$\text{ELBO}(Q) \coloneqq E_Q[\log P(\mathbf{u}, \mathbf{f}, \mathbf{m}|\{Y_b\}; \mathbf{x}) - \log Q(\mathbf{u}, \mathbf{f}, \mathbf{m})],$$

as $KL(Q||P(\mathbf{u}, \mathbf{f}, \mathbf{m}|\{Y_b\}; \mathbf{x})) = -\text{ELBO}(Q) + \log P(\{Y_b\}; \mathbf{x})$. Based on standard VI theory, if $Q \in \text{argmax}_{Q \in \mathcal{Q}} \text{ELBO}(Q)$, then Q has the following form:

$$\begin{aligned} Q(\mathbf{u}, \mathbf{f}, \mathbf{m}) &= Q(\mathbf{u})P(\mathbf{f}|\mathbf{u})\Pi_{b \in \mathcal{B}} Q(M_b), \text{ with} \quad (14) \\ Q(\mathbf{u}) &= \mathcal{N}(\mathbf{u}; \mu^u, \Sigma^u) \quad (15) \\ Q(M_b) &\propto P(Y_b|M_b)\Pi_{i \in b}\mathcal{N}(m_i; \mu_i^M, 1) \text{ for } b \in \mathcal{B}. \quad (16) \end{aligned}$$

Here, $\mu^u \in \mathbb{R}^R$ and $\Sigma^u \in \mathbb{R}^{R \times R}$ specifying the mean and covariance of $Q(\mathbf{u})$ and $\mu^M = (\mu_1^M, \ldots, \mu_N^M)$, with $\mu_i^M \in \mathbb{R}$ for $i \in [N]$, are the variational parameters parametrizing Q.

Given this parametrization of Q, standard mean field theory (see Appendix) gives closed form updates to the variational parameters to maximize $\text{ELBO}(Q)$:

$$\begin{aligned} \Sigma^u &\leftarrow (K_{\mathbf{zz}}^{-1} + K_{\mathbf{zz}}^{-1} K_{\mathbf{zx}} K_{\mathbf{xz}} K_{\mathbf{zz}}^{-1})^{-1} \quad (17) \\ \mu^u &\leftarrow \Sigma^u K_{\mathbf{zz}}^{-1} K_{\mathbf{zx}} E_{Q(\mathbf{m})}[\mathbf{m}] \quad (18) \\ \mu^M &\leftarrow K_{\mathbf{xz}} K_{\mathbf{zz}}^{-1} \mu^u \quad (19) \end{aligned}$$

In Equation 18, $E_{Q(\mathbf{m})}[\mathbf{m}]$ can be computed by separately computing $E_{Q(M_b)}[M_b]$ for each $b \in \mathcal{B}$ in closed form. For a given bag b, there are two cases, depending on whether it is a positive or negative bag.

If $Y_b = 0$, then combining Equations 11 and 16, we get $Q(M_b) \propto \Pi_{i \in b} 1[0 > m_i]\mathcal{N}(m_i; \mu_i^M, 1)$, so that the $\{m_i\}_{i \in b}$ (i.e. the entries of M_b) are independent of each other under $Q(M_b)$. Thus for $i \in b$,

$$Q(m_i) \propto 1[0 > m_i]\mathcal{N}(m_i; \mu_i^M, 1). \quad (20)$$

Thus if $Y_b = 0$, for $i \in b$, $Q(m_i)$ is a *truncated* normal distribution, so that

$$\begin{aligned} E_{Q(m_i)}[m_i] &= S(\mu_i^M), \text{ where} \quad (21) \\ S(m) &\coloneqq m - \frac{\mathcal{N}(m; 0, 1)}{1 - \text{PROBIT}^{-1}(m)} \quad (22) \end{aligned}$$

is the mean of a $\mathcal{N}(m, 1)$ distribution right-truncated at 0.

If $Y_b = 1$, then combining Equations 12 and 16, we get

$$Q(M_b) = \tfrac{1}{Z}\big(1 - \Pi_{i \in b} 1[0 > m_i]\big)\Pi_{i \in b}\mathcal{N}(m_i; \mu_i^M, 1), \quad (23)$$

where normalizing constant Z is readily calculated as

$$\begin{aligned} Z &= 1 - \int_{M_b} \Pi_{i \in b} 1[0 > m_i]\mathcal{N}(m_i; \mu_i^M, 1) dM_b \\ &= 1 - \Pi_{i \in b} \int_{m_i} 1[0 > m_i]\mathcal{N}(m_i; \mu_i^M, 1) dm_i \\ &= 1 - \Pi_{i \in b}(1 - \text{PROBIT}^{-1}(\mu_i^M)). \end{aligned}$$

Thus, if $Y_b = 1$, for $i \in b$,

$$\begin{aligned} &E_{Q(m_i)}[m_i] \\ &= \tfrac{1}{Z} \int_{M_b} m_i\big(1 - \Pi_{i \in b} 1[0 > m_i]\big)\Pi_{i \in b}\mathcal{N}(m_i; \mu_i^M, 1) dM_b \\ &= \tfrac{1}{Z}(\mu_i^M - \int_{M_b} m_i \Pi_{i \in b} 1[0 > m_i]\mathcal{N}(m_i; \mu_i^M, 1) dM_b) \\ &= \tfrac{1}{Z}(\mu_i^M - S(\mu_i^M)\Pi_{i \in b}(1 - \text{PROBIT}^{-1}(\mu_i^M))). \end{aligned}$$

3.3 MIL-GP-Probit to MIL-GP-Logistic Comparison

The MIL-GP-Logistic model of [10] is nearly identical to our MIL-GP-Probit model, except that they use a logit link to relate instance scalars to instance label probabilities, whereas we use a probit link. In more detail, the MIL-GP-Probit model is given by Equations 5, 6, 7, 8, whereas the MIL-GP-Logistic model is given by Equations 5, 6, 4, 8. As the logit and probit link are known to give similar predictive performance [11] in Bayesian regression models, we can conclude that the MIL-GP-Logistic and MIL-GP-Probit models should also give similar predictive performance, *provided* accurate posterior inference can be performed.

However, the inference procedure for MIL-GP-Logistic makes an approximation that we do not make, which

as our experiments will show, hurt their predictive performance. Recall to approximate the posterior of the MIL-GP-Probit model, we find $\text{argmax}_{Q\in\mathcal{Q}} \text{ELBO}(Q)$, where $\text{ELBO}(\cdot)$ is given in Section 3.2, and the variational family $\mathcal{Q}$ is given by Equation 13. The inference procedure for the MIL-GP-Logistic model infers the posterior over their model's parameters, $P(\mathbf{u}, \mathbf{f}, \mathbf{h}|\{Y_b\}; \mathbf{x})$ by approximating it with a variational distribution $Q(\mathbf{u}, \mathbf{f}, \mathbf{h}) \in \bar{\mathcal{Q}}$, where $\bar{\mathcal{Q}}$ are distributions factoring as $Q(\mathbf{u})Q(\mathbf{h})P(\mathbf{f}|\mathbf{u}; \mathbf{x})$, and $\mathbf{h} \coloneqq (h_1, \ldots, h_N)$. This factorization assumption is similar to the one we make.

Where the inference methods for the two models differ is the variational objective that is optimized. Whereas we maximize the standard ELBO objective *directly* over the variational family, they maximize a *lower bound* on the ELBO over the variational family. In particular, for the MIL-GP-Logistic model, they find $\text{argmax}_{Q\in\bar{\mathcal{Q}}} \text{ELBO}^-(Q)$, where $\text{ELBO}^-(Q) \leq \text{ELBO}(Q) \coloneqq E_Q[\log P(\mathbf{u}, \mathbf{f}, \mathbf{h}|\{Y_b\}; \mathbf{x}) - \log Q(\mathbf{u}, \mathbf{f}, \mathbf{h})]$.

Please see [10] for the exact form of $\text{ELBO}^-(Q)$. What matters is that above inequality is in general not an equality. As maximizing the lower bound of a function is suboptimal, their inference method can and we show, does lead to lowered predictive performance. The reason they cannot maximize the ELBO directly is because they use the logit link, and are forced to use the Jaakkola lower bound. On the other hand, our use of the probit link along with the variable augmentation approach allows us to maximize the ELBO directly.

3.4 Making Test Predictions

For a test bag b^* with instance features $X_{b^*} \coloneqq \{x_i^*\}_{i\in b^*}$, the posterior predictive distribution over $\mathbf{u}$, instance scalars $F_{b^*} \coloneqq \{f_i^*\}_{i\in b^*}$, augmenting variables $M_{b^*} \coloneqq \{m_i^*\}_{i\in b^*}$, and bag label Y_{b^*} is approximated analogous to the augmented MIL-GP-Probit model of Section 3.2:

$$
\begin{aligned}
&Q(\mathbf{u}, F_{b^*}, M_{b^*}, Y_{b^*}) && (24)\\
&= Q(\mathbf{u})P(F_{b^*}|\mathbf{u})P(M_{b^*}|F_{b^*})P(Y_{b^*}|M_{b^*})\\
&\approx P(\mathbf{u}, F_{b^*}, M_{b^*}, Y_{b^*}|\{Y_b\}; \{X_b\}, \mathbf{z}, \theta), && (25)
\end{aligned}
$$

where $F_{b^*}|\mathbf{u}; X_{b^*} \sim \mathcal{N}(K_{X_{b^*}\mathbf{z}}K_{\mathbf{zz}}^{-1}\mathbf{u}, \text{diag}(\mathcal{K}^*))$, with $\mathcal{K}^* \coloneqq K_{X_{b^*}X_{b^*}} - K_{X_{b^*}\mathbf{z}}K_{\mathbf{zz}}^{-1}K_{\mathbf{z}X_{b^*}}$, $m_i^* \sim \mathcal{N}(f_i^*, 1)$, $h_i^* \sim 1[m_i^* > 0]$ for $i \in b^*$, and following Equation 12, $P(Y_{b^*} = 1|M_{b^*}) = 1 - \Pi_{i\in b^*}1[0 > m_i^*]$. Note that $h_i^*|m_i^* \sim 1[m_i^* > 0]$ for $i \in b^*$, giving the posterior predictive distribution for each test instance label h_i^*. Marginalizing out $\mathbf{u}$ gives $Q(F_{b^*}) = \mathcal{N}(\mu^{F_{b^*}}, \Sigma^{F_{b^*}})$, where $\mu^{F_{b^*}}, \Sigma^{F_{b^*}}$ are given by the standard Gaussian convolution formula. Then, $Q(M_{b^*}) = \mathcal{N}(\mu^{f_{b^*}}, \Sigma^{f_{b^*}} + I_{|b^*|})$, where $I_{|b^*|}$ is the $|b^*| \times |b^*|$ identity matrix. The Gaussian characterization of $Q(M_{b^*})$ makes it possible to make both test instance and bag label predictions.

To make instance label predictions, note for $i \in b^*$,

$$
\begin{aligned}
Q(h_i^* = 1) &= \int_{m_i^*} Q(m_i^*)P(h_i^*|m_i^*)dm_i^*\\
&= \int_{m_i^*} \mathcal{N}(m_i^*; \mu_{ii}^{F_{b^*}}, \Sigma_{ii}^{F_{b^*}} + 1)1[m_i^* > 0]dm_i^*\\
&= \text{PROBIT}^{-1}\left(\frac{\mu_i^{F_{b^*}}}{\Sigma_{ii}^{F_{b^*}} + 1}\right).
\end{aligned}
$$

To make bag label predictions, note

$$
\begin{aligned}
Q(Y_{b^*} = 1) &= \int_{M_b^*} Q(M_b^*)P(Y_{b^*} = 1|M_{b^*})dM_b^*\\
&= 1 - \int_{M_b^*} \mathcal{N}(\mu^{F_{b^*}}, \Sigma^{F_{b^*}} + I_{|b^*|})\Pi_{i\in b^*}1[0 > m_i^*]dM_b^*.
\end{aligned}
$$

The above integral is a Gaussian integral subject to linear inequality constraints, and can be calculated efficiently using the method of [6]. Note that the bag label prediction accounts for *dependence* between the instance labels in the bag, as $\Sigma^{F_{b^*}}$ is not diagonal. This is the philosophically correct way to make bag label predictions. Although we do not target bag label predictions in this work, having *some* mechanism for making them would still be important if performing active learning, where uncertainty in bag label predictions might be used to select the bag labels to acquire.

Note that to make bag label predictions with the MIL-GP-Logistic model of [10], one would need to assume instance labels are independent of each other, due to the intractability of the logit link. This is philosophically incorrect simply due to the model not asserting instance label independence. It is also intuitively incorrect; suppose a bag contained many identical instances. Assuming independence of instance labels implies the bag has almost no chance of being negative, as the probability all the instance labels are independently negative is vanishingly low.

3.5 Gibbs Sampling

We also derive a Gibbs sampling scheme for our MIL-GP-Probit model. To the best of our knowledge, this is the first (asymptotically) exact posterior inference scheme for *any* Bayesian model for the MIL scenario, and provides us not only ground truth, but the ability to assess the effect of approximate variational inference.

We derive Gibbs sampling on the *augmented* MIL-GP-Probit model of Section 3.2, for which the full conditional distributions of $P(\mathbf{u}, \mathbf{f}, \mathbf{m}|\{Y_b\}; \mathbf{x})$ can be derived. First, we integrate out $\mathbf{f}$ to obtain $P(\mathbf{u}, \mathbf{m}|\{Y_b\}; \mathbf{x})$. It is then straightforward (see Appendix) to derive the full conditional distributions of $\mathbf{u}$ and each m_i for $i \in [N]$, so that *collapsed* Gibbs sampling can be performed:

$$P(\mathbf{u}|\mathbf{m}, \{Y_b\}; \mathbf{x}) = \mathcal{N}(\mathbf{u}; \Sigma^u K_{\mathbf{zz}}^{-1}K_{\mathbf{zx}}\mathbf{m}, \Sigma^u),$$

where Σ^u is given by Equation 17. For the full conditional distribution of m_i for $i \in b$, there are 3 cases depending on whether bag b is a negative bag, and if not, whether any other instance within the bag are positive given the current parameter values:

$$P(m_i|\{m_{i'}\}_{i' \in b, i' \neq i}, \mathbf{u}, \{Y_b\}; \mathbf{x})$$

$$\begin{cases} \propto \mathcal{N}(m_i; \mu_i^M, 1)1[m_i < 0] \text{ if } Y_b = 0 & (26) \\ \propto \mathcal{N}(m_i; \mu_i^M, 1)1[m_i > 0] & (27) \\ \quad \text{if } Y_b = 1 \text{ and } m_{i'} < 0 \; \forall \; i' \in b, i' \neq i & \\ = \mathcal{N}(m_i; \mu_i^M, 1) \text{ otherwise,} & (28) \end{cases}$$

where $\mu^M := (\mu_1^M, \ldots, \mu_N^M) := K_{\mathbf{xz}} K_{\mathbf{zz}}^{-1} \mathbf{u}$. The conditional distributions of Equations 26 and 27 are univariate *truncated* normal distributions, which can be efficiently sampled from.

4 Experiments

Firstly, our MIL-GP-Probit model admits a variational inference procedure that directly maximizes the ELBO. On the other hand, MIL-GP-Logistic model of [10] is similar to ours, but as described in Section 3.3, uses a variational inference method that maximizes a *lower bound* on the ELBO, instead of maximizing it directly. Although their method was shown already to be the state of the art, this approximation potentially hurts their predictive performance. Thus the first question we answer with experiments is:

Q1: *How does the predictive performance of our* MIL-GP-Probit *model compare to that of the* MIL-GP-Logistic *model of [10]?*

Secondly, we have also developed the first (asymptotically) exact posterior inference procedure for *any* Bayesian model for the MIL scenario. In particular, we have developed a Gibbs sampler for our MIL-GP-Probit model. Although MCMC procedures are slower, they give us a "ground truth". Thus the second question we answer with our experiments is:

Q2: *How does the predictive performance of our* MIL-GP-Probit *model differ when (exact) posterior inference via Gibbs sampling is used instead of variational inference?*

To answer these questions, we evaluated the following methods:

- MIL-GP-Probit: Our MIL-GP-Probit model, with the variational inference method of Section 3.2 running for 25 iterations.
- MIL-GP-Probit-Gibbs: Our MIL-GP-Probit model, with the Gibbs sampling inference method of Section 3.5 providing 5000 samples with no thinning, 1000 samples of burn-in.
- MIL-GP-Logistic: The MIL-GP-Logistic model of [10], with their variational inference method running for 25 iterations (as they did).
- MIL-GP-Logistic-LM: [10] also developed a "large-margin" extension of their MIL-GP-Logistic model, which encourages the decision boundary to lie far away from instances; see their paper for details. This is that method, with the default hyperparameters of $C = 2, V = 2$.

For all methods, we chose $R = 50$ inducing points via K-means-++[3]. We only compare our methods to the methods of [10], as the latter were already shown to be the state of the art compared to other methods. Please see [10] for comparison of their methods with others.

Since this paper tackles the *instance* label prediction problem, we assume that given a test instance x^*, a method can produce p^*, the predicted probability the instance label is positive. For a given set of test instances, the following criteria are used to evaluate the predictions of the test instance labels, $\{p^*\}$, relative to their true labels, $\{h^*\}$:

- AUC: Area under the Receiver Operating Curve.
- Loglik: The predictive log-likelihood over the test set of instances: $\frac{1}{N^*} \sum h^* \log p^* + (1 - h^*) \log(1 - p^*)$, where N^* is the number of test instances. This is the standard criteria for evaluating probabilistic models, measuring the KL-divergence between the predicted test instance label distribution and the true test instance label distribution.
- MAP: Mean-Average-Precision, i.e. the area under the precision-recall curve.

4.1 Experiments on 20 Newsgroups datasets

The 20 Newsgroups dataset is a collection of 20 MIL datasets first introduced by [25], where each dataset is associated with a single newsgroup (out of 20), and consists of bags of around 40 instances. Each instance is positive if it came from that single newsgroup, and is negative otherwise. Instances are represented by 200 TF-IDF features. Although it is synthetic, it has been widely used as a benchmark to evaluate instance label predictive performance for the MIL scenario, because only about 3% of the instances in each positive bag are positive instances.

For each of the 20 datasets, we perform 10 rounds of 10-fold cross-validation, using the publicly available folds from [25]; a training fold contains labelled bags, and a test fold contains test instances whose label is to be predicted. Following [10], for all evaluated methods, we use the Gaussian kernel, with the lengthscale set to the square root of the instance feature dimension, and

Criteria	AUC		Loglik		MAP	
Method / dataset	MIL-GP-Probit	MIL-GP-Logistic	MIL-GP-Probit	MIL-GP-Logistic	MIL-GP-Probit	MIL-GP-Logistic
atheism	0.969	**0.974**	**-0.036**	-0.158	**0.714**	0.700
graphics	0.901	**0.928**	**-0.052**	-0.164	**0.796**	0.787
windows	0.903	**0.922**	**-0.036**	-0.159	**0.543**	0.541
pc	0.909	**0.955**	**-0.038**	-0.156	**0.708**	0.700
mac	0.943	**0.947**	**-0.042**	-0.159	0.761	**0.763**
windows.x	0.946	**0.972**	**-0.056**	-0.168	0.734	**0.736**
forsale	0.908	**0.945**	**-0.034**	-0.156	0.521	**0.526**
rec.autos	**0.944**	0.935	**-0.051**	-0.170	**0.746**	0.741
motorcycles	0.979	**0.981**	**-0.040**	-0.169	0.685	**0.720**
baseball	0.945	**0.976**	**-0.051**	-0.174	0.759	**0.776**
hockey	0.988	**0.990**	**-0.075**	-0.181	0.914	**0.923**
sci.crypt	0.988	**0.995**	**-0.042**	-0.161	0.703	**0.773**
electronics	**0.990**	0.967	**-0.048**	-0.154	**0.926**	0.918
sci.med	**0.956**	0.951	**-0.054**	-0.171	**0.760**	0.742
sci.space	**0.962**	0.981	**-0.049**	-0.175	0.731	**0.752**
christian	0.960	**0.971**	**-0.040**	-0.178	0.747	**0.750**
guns	**0.979**	0.975	**-0.048**	-0.163	0.702	**0.723**
mideast	0.974	**0.974**	**-0.050**	-0.160	0.805	**0.850**
politics	0.966	**0.969**	**-0.037**	-0.153	0.637	**0.646**
religion	0.932	**0.937**	**-0.038**	-0.171	**0.561**	0.531
wins	5	15	20	0	8	12

Table 2: On the 20 newsgroups dataset collection, our method (MIL-GP-Probit) has higher predictive log-likelihood than MIL-GP-Logistic [10] on all 20 datasets, and is comparable in terms of AUC and MAP.

use kernel PCA to reduce the instance feature dimension to 100. Table 2 shows the predictive performance of MIL-GP-Probit compared to MIL-GP-Logistic on each of the 20 MIL datasets in the 20 Newsgroups dataset. The two methods are comparable in terms of AUC and MAP. However, MIL-GP-Probit has higher predictive log-likelihood than MIL-GP-Logistic on *all 20 datasets*. Since the MIL-GP-Probit and MIL-GP-Logistic *models* are very similar (see Section 3.3), one can conclude the reason the MIL-GP-Probit outperforms the MIL-GP-Logistic *method* in terms of predictive log-likelihood is because the former uses a variational inference method that *directly* optimizes the ELBO, whereas the latter uses a variational inference method that optimizes a *lower bound* on the ELBO. Table 3 summarizes the performance of all considered methods across the 20 datasets by giving the average value of each criteria across all 20 datasets for each method. We once again see that MIL-GP-Probit is comparable to MIL-GP-Logistic in terms of AUC and MAP. Comparing MIL-GP-Probit to MIL-GP-Probit-Gibbs in terms of predictive log-likelihood, we see that the former, which uses a variational approximation to the posterior, is actually comparable to the latter, which avoids that approximation by using Gibbs sampling. In fact, the AUC and MAP of MIL-GP-Probit-Gibbs are lower; this is due to a peculiarity of this particular dataset collection: that the true posterior predictive probabilities under the assumed model are all close to 0.5. This causes all posterior sampling methods to do poorly on evaluation metrics based on correctly ordering the instances by true posterior predictive probability; we now elaborate.

For any posterior sampler, the estimators of posterior predictive probabilities are Monte Carlo means, and are used to calculate AUC (and MAP). Suppose true negatives and true positives have a true posterior predictive probability of 0.49 and 0.51, respectively (i.e. all close to 0.5). Due to estimator variance, the estimated posterior predictive probabilities for some true negative instances would be higher than those of some true positive instances. As instances are incorrectly ordered, this lowers AUC. This happens even if the sampler were "perfect" (directly samples the true posterior) and the model were "perfect" (true posterior predictive probabilities would give a perfect AUC). Under our model, the true posterior predictive probabilities for the 20 Newsgroups dataset collection are all close to 0.5, and so our Gibbs sampler has poor AUC simply because it is *a* sampler. We also note that our model is "correct" in the sense that the dataset was intentionally created to make it hard to distinguish positive from negative instances (by making positive instances a very small fraction of the instances in positive bags), so that the posterior predictive probabilities should all be close to 0.5. The AUC can be improved simply by obtaining more posterior samples to lower estimator variance.

	AUC	Loglik	MAP
MIL-GP-Probit-Gibbs	0.853	-0.045	0.651
MIL-GP-Probit	0.952	-0.046	0.723
MIL-GP-Logistic	0.962	-0.165	0.730
MIL-GP-Logistic-LM	0.957	-0.343	0.725

Table 3: The criteria for all considered methods, averaged over all 20 datasets of the 20 Newsgroups dataset collection. MIL-GP-Probit (ours) has much higher predictive log-likelihood than MIL-GP-Logistic ([10]), and is comparable in terms of AUC and MAP. Despite its variational approximation, MIL-GP-Probit has comparable predictive log-likelihood to our exact inference method MIL-GP-Probit-Gibbs, whose AUC and MAP suffer due to the difficulty of distinguishing between instances with similar true probabilities via sampling. MIL-GP-Logistic-LM does poorly.

4.2 Experiments on additional non-synthetic datasets

We further evaluated all considered methods on an additional 59 non-synthetic MIL datasets. [18] introduces several datasets for the *multi-label* multiple instance (MIML) learning scenario, where each instance is associated with one of $K > 2$ possible classes, and the label for a bag indicates the union of the instance labels within it. For a given MIML dataset where each instance is one of K classes, we can transform it into K separate MIL datasets by treating each class in turn as the positive class.

Criteria	AUC		Loglik		MAP	
Method / dataset	MIL-GP-Probit	MIL-GP-Logistic	MIL-GP-Probit	MIL-GP-Logistic	MIL-GP-Probit	MIL-GP-Logistic
Salad_0	**0.852**	0.831	**-0.252**	-0.258	**0.530**	0.477
Salad_1	**0.842**	0.838	-0.258	**-0.255**	0.483	**0.487**
Salad_2	**0.859**	0.800	**-0.166**	-0.191	**0.483**	0.363
Salad_3	0.874	**0.878**	-0.412	**-0.287**	0.695	**0.771**
Salad_4	**0.993**	0.991	**-0.106**	-0.143	0.969	**0.974**
Salad_5	0.825	**0.835**	-0.198	**-0.185**	0.372	**0.445**
Voc12_0	**0.643**	0.542	-0.025	**-0.025**	**0.012**	0.009
Voc12_1	**0.662**	0.534	**-0.073**	-0.075	**0.056**	0.036
Voc12_10	**0.551**	0.446	**-0.083**	-0.091	**0.044**	0.034
Voc12_11	**0.751**	0.677	**-0.061**	-0.063	**0.078**	0.043
Voc12_12	**0.817**	0.746	**-0.051**	-0.053	**0.092**	0.047
Voc12_13	**0.812**	0.739	**-0.068**	-0.072	**0.139**	0.086
Voc12_14	**0.724**	0.714	**-0.304**	-0.346	**0.534**	0.518
Voc12_15	**0.743**	0.694	**-0.086**	-0.090	**0.123**	0.090
Voc12_16	**0.708**	0.576	-0.034	**-0.034**	**0.023**	0.015
Voc12_17	0.627	**0.627**	**-0.083**	-0.085	**0.062**	0.056
Voc12_18	**0.552**	0.385	-0.028	**-0.027**	**0.010**	0.007
Voc12_19	**0.861**	0.816	**-0.063**	-0.068	**0.159**	0.100
Voc12_2	**0.727**	0.696	-0.024	**-0.023**	**0.017**	0.014
Voc12_3	**0.714**	0.686	-0.041	**-0.041**	**0.035**	0.031
Voc12_4	**0.554**	0.528	-0.114	**-0.114**	**0.077**	0.072
Voc12_5	**0.855**	0.805	**-0.058**	-0.064	**0.106**	0.072
Voc12_6	**0.876**	0.853	**-0.117**	-0.129	**0.359**	0.285
Voc12_7	**0.693**	0.549	-0.042	**-0.041**	**0.027**	0.017
Voc12_8	**0.808**	0.807	-0.169	-0.174	**0.375**	0.357
Voc12_9	**0.652**	0.605	-0.051	-0.052	**0.030**	0.026
hja_0	**0.756**	0.734	**-0.108**	-0.159	**0.134**	0.117
hja_1	0.645	**0.694**	-0.155	**-0.129**	0.271	**0.292**
hja_10	**0.837**	0.638	**-0.044**	-0.055	**0.177**	0.040
hja_11	**0.936**	0.932	-0.158	**-0.101**	0.874	**0.875**
hja_12	**0.763**	0.746	**-0.032**	-0.037	**0.025**	0.023
hja_2	**0.843**	0.817	**-0.095**	-0.123	0.203	**0.279**
hja_3	**0.934**	0.927	-0.088	**-0.088**	0.490	**0.502**
hja_4	0.568	**0.635**	**-0.023**	-0.028	0.009	**0.011**
hja_5	0.759	**0.788**	**-0.060**	-0.082	0.058	**0.070**
hja_6	0.597	**0.700**	**-0.012**	-0.015	0.005	**0.010**
hja_7	**0.882**	0.786	**-0.066**	-0.077	**0.399**	0.239
hja_8	**0.977**	0.944	**-0.028**	-0.050	**0.715**	0.473
hja_9	0.619	**0.661**	**-0.031**	-0.043	0.015	**0.018**
msrcv2_0	**0.843**	0.768	**-0.160**	-0.184	**0.415**	0.246
msrcv2_1	**0.873**	0.870	**-0.181**	-0.215	**0.694**	0.614
msrcv2_10	**0.928**	0.896	**-0.083**	-0.094	**0.327**	0.210
msrcv2_11	**0.792**	0.575	**-0.067**	-0.071	**0.079**	0.040
msrcv2_12	**0.632**	0.327	-0.050	**-0.049**	**0.029**	0.018
msrcv2_13	**0.852**	0.821	**-0.030**	-0.032	**0.312**	0.207
msrcv2_14	**0.772**	0.633	-0.036	**-0.035**	**0.037**	0.025
msrcv2_15	**0.564**	0.405	**-0.074**	-0.075	**0.043**	0.035
msrcv2_17	**0.581**	0.528	-0.034	**-0.032**	0.033	**0.035**
msrcv2_18	**0.833**	0.754	**-0.155**	-0.169	**0.403**	0.225
msrcv2_2	**0.827**	0.726	**-0.169**	-0.188	**0.309**	0.224
msrcv2_20	**0.687**	0.519	-0.041	**-0.040**	**0.024**	0.015
msrcv2_21	**0.762**	0.726	**-0.099**	-0.105	**0.142**	0.137
msrcv2_22	**0.754**	0.691	**-0.066**	-0.070	**0.077**	0.062
msrcv2_3	**0.828**	0.753	**-0.077**	-0.082	**0.113**	0.075
msrcv2_5	**0.815**	0.654	**-0.055**	-0.056	**0.063**	0.032
msrcv2_6	**0.881**	0.842	**-0.154**	-0.192	**0.522**	0.385
msrcv2_7	**0.812**	0.787	**-0.043**	-0.048	**0.053**	0.044
msrcv2_8	**0.826**	0.753	**-0.049**	-0.054	**0.056**	0.039
msrcv2_9	**0.889**	0.860	**-0.084**	-0.095	**0.305**	0.187
wins	51	8	42	17	46	13

Table 4: MIL-GP-Probit (ours) is better than MIL-GP-Logistic ([10]) on the majority of 59 non-synthetic MIL datasets across all criteria.

To generate the 59 MIL datasets, we took all 4 non-synthetic MIML datasets of [18] and applied the above transformation. Each resulting MIL dataset is denoted by the name of the original MIML dataset and the index of the positive class. Those datasets are:

- msrcv2: Image annotation dataset where each bag is an image and patches within an image are the instances. Features are a 48-dimensional histogram of gradients and colors vector. 1636 instances, 469 bags, 23 classes, 48 features, 2.5 mean bag label size.
- Voc12: Image annotation dataset like msrcv2. Same features as msrcv2. 4142 instances, 1053 bags, 20 classes, 48 features, 2.3 mean bag label size.
- Salad: 100-second time windows are bags divided into 2 second instances. Each instance's label is the activity a human subject was doing while in a kitchen, such as cutting cheese, cutting lettuce, mixing ingredients. Features are derived from accelerometers worn on the subject's wrist. 2020 instances, 124 bags, 6 classes, 58 features, 2.3 mean bag label size.
- hja: Birdsong spectrogram dataset where each bag is a 10 second time window, and instances are patches of high intensity within it obtained using [17]. The instance label is the bird species the patch came from. Instances with no ground truth are removed. 4983 instances, 533 bags, 13 classes, 39 features, 2.1 mean bag label size.

For each of the 59 datasets, we performed 100 rounds of cross-validation, where in each round, we randomly select $\frac{1}{2}$ of the bags for training data, and the instances in the remaining bags as the test instances. As before, we follow the heuristic of [10] and set the length scale of the kernel to the square root of the instance feature dimension. However, for these 59 datasets we use the exponential kernel, which modeled the data better.

Table 4 shows that MIL-GP-Probit does better than MIL-GP-Logistic on 51, 47, and 46 of the 59 datasets in terms of AUC, predictive log-likelihood, and MAP, respectively. MIL-GP-Probit turns out to have similar performance to MIL-GP-Probit-Gibbs: Table 5 shows that the average of the 3 criteria across all 59 datasets is similar for the two methods; see the Appendix for results for individual datasets. Thus MIL-GP-Probit is comparable in performance to MIL-GP-Probit-Gibbs, which performs (asymptotically) exact inference, but takes longer to run. Finally, MIL-GP-Logistic, as Table 4 already indicated, has worse average performance than our methods, and MIL-GP-Logistic-LM does worse.

	AUC	Loglik	MAP
MIL-GP-Probit-Gibbs	0.764	-0.097	0.227
MIL-GP-Probit	0.770	-0.094	0.225
MIL-GP-Logistic	0.713	-0.099	0.190
MIL-GP-Logistic-LM	0.539	-0.133	0.110

Table 5: The criteria for all considered methods, averaged over all 59 non-synthetic MIL datasets. Our methods (MIL-GP-Probit, MIL-GP-Probit-Gibbs), do better than those of [10] (MIL-GP-Logistic, MIL-GP-Logistic-LM), in terms of AUC and MAP. Despite its variational approximation, MIL-GP-Probit has comparable performance to MIL-GP-Probit-Gibbs, which characterizes the exact posterior via sampling, but is slower. MIL-GP-Logistic-LM does poorly.

5 Related Work and Conclusion

We have created MIL-GP-Probit, a Bayesian Gaussian Process model for the Multiple Instance Learning scenario for tackling the instance label prediction problem. Our model is an adaptation of the MIL-GP-Logistic model of [10] to use a probit instead of logit link function. This seemingly small change enables our first contribution: the variational inference procedure we provide. Whereas the inference procedure of [10] maximizes a lower bound on the ELBO objective, our inference procedure leverages the analytical tractability of the probit link to maximize the ELBO directly, and achieves higher predictive performance than [10] on the benchmark 20 Newsgroups dataset and 59 additional datasets. Secondly, we develop a Gibbs posterior sampling scheme for our model, which enables (asymptotically) exact posterior inference. To the best of our knowledge, this is the first MCMC posterior inference procedure for *any* Bayesian model for the MIL scenario.

Related work includes aforementioned work on multiple instance learning. Our work is also related to probabilistic methods for learning when training data is not ideal, such as when given aggregate counts [16], or group proportions [15]. Use of the probit link in probabilistic models has been shown to enable efficient and accurate posterior inference in Bayesian linear regression [1] and multi-class Gaussian Process models [7]. Our use of the probit link in the Bayesian MIL scenario can be seen as an application of that insight.

Future work includes allowing extensions of the standard MIL bag label assumption as in [9], incorporating a feature transformation step as in [23], and using our model in Bayesian Active Learning [12]. In fact, one of the advantages of our model is that computing the BALD [12] acquisition function is tractable for it, unlike for the MIL-GP-Logistic model of [10].

6 Acknowledgements

This work was supported by the Laboratory Directed Research and Development (LDRD) program at Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

[1] James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

[2] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*, pages 577–584, 2003.

[3] David Arthur and Sergei Vassilvitskii. K-means++: the advantages of careful seeding. In *In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.

[4] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[5] James Foulds and Padhraic Smyth. Multi-instance mixture models and semi-supervised learning. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 606–617. SIAM, 2011.

[6] Alan Genz and Frank Bretz. *Computation of multivariate normal and t probabilities*, volume 195. Springer Science & Business Media, 2009.

[7] Mark Girolami and Simon Rogers. Variational bayesian multinomial probit regression with gaussian process priors. *Neural Computation*, 18(8):1790–1817, 2006.

[8] Xinze Guan, Raviv Raich, and Weng-Keen Wong. Efficient multi-instance learning for activity recognition from time series data using an autoregressive hidden markov model. In *International Conference on Machine Learning*, pages 2330–2339, 2016.

[9] Hossein Hajimirsadeghi, Wang Yan, Arash Vahdat, and Greg Mori. Visual recognition by counting instances: A multi-instance cardinality potential kernel. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2596–2605, 2015.

[10] Manuel Haußmann, Fred A Hamprecht, and Melih Kandemir. Variational bayesian multiple instance learning with gaussian processes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6570–6579, 2017.

[11] Chris C Holmes, Leonhard Held, et al. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian analysis*, 1(1):145–168, 2006.

[12] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[13] Melih Kandemir and Fred A Hamprecht. Instance label prediction by dirichlet process multiple instance learning. In *UAI*, pages 380–389, 2014.

[14] Minyoung Kim and Fernando De la Torre. Gaussian processes multiple instance learning. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 535–542, 2010.

[15] Hendrik Kück and Nando de Freitas. Learning about individuals from group statistics. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 332–339, 2005.

[16] Ho Chung Law, Dino Sejdinovic, Ewan Cameron, Tim Lucas, Seth Flaxman, Katherine Battle, and Kenji Fukumizu. Variational learning on aggregate outputs with gaussian processes. In *Advances in Neural Information Processing Systems*, pages 6081–6091, 2018.

[17] Lawrence Neal, Forrest Briggs, Raviv Raich, and Xiaoli Z Fern. Time-frequency segmentation of bird song in noisy acoustic environments. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2012–2015. IEEE, 2011.

[18] Anh T Pham, Raviv Raich, and Xiaoli Z Fern. Dynamic programming for instance annotation in multi-instance multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2381–2394, 2017.

[19] Soumya Ray and Mark Craven. Supervised versus multiple instance learning: An empirical comparison. In *Proceedings of the 22nd international conference on Machine learning*, pages 697–704, 2005.

[20] Vikas C Raykar, Balaji Krishnapuram, Jinbo Bi, Murat Dundar, and R Bharat Rao. Bayesian multiple instance learning: automatic feature selection and inductive transfer. In *Proceedings of the 25th international conference on Machine learning*, pages 808–815, 2008.

[21] Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. *Advances in neural information processing systems*, 18:1257–1264, 2005.

[22] Sudheendra Vijayanarasimhan and Kristen Grauman. Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[23] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378, 2016.

[24] Cha Zhang, John C Platt, and Paul A Viola. Multiple instance boosting for object detection. In *Advances in neural information processing systems*, pages 1417–1424, 2006.

[25] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th annual international conference on machine learning*, pages 1249–1256. ACM, 2009.

[26] Zhi-Hua Zhou and Jun-Ming Xu. On the relation between multi-instance learning and semi-supervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1167–1174, 2007.