# Variational Weakly Supervised Gaussian Processes

Melih Kandemir[1]
melih.kandemir@iwr.uni-heidelberg.de
Manuel Haußmann[1]
manuel.haussmann@iwr.uni-heidelberg.de
Ferran Diego[1]
ferran.diego@iwr.uni-heidelberg.de
Kumar Rajamani[2]
KumarThirunellai.Rajamani@in.bosch.com
Jeroen van der Laak[3]
Jeroen.vanderLaak@radboudumc.nl
Fred A. Hamprecht[1]
fred.hamprecht@iwr.uni-heidelberg.de

[1] Heidelberg University, HCI
Heidelberg, Germany
[2] Robert Bosch Engineering
Bangalore, India
[3] Radboud University Medical Center
Nijmegen, Netherlands

**Abstract**

We introduce the first model to perform weakly supervised learning with Gaussian processes on up to millions of instances. The key ingredient to achieve this scalability is to replace the standard assumption of MIL that the bag-level prediction is the maximum of instance-level estimates with the accumulated evidence of instances within a bag. This enables us to devise a novel variational inference scheme that operates solely by closed-form updates. Keeping all its parameters but one fixed, our model updates the remaining parameter to the global optimum. This virtue leads to charmingly fast convergence, fitting perfectly to large-scale learning setups. Our model performs significantly better in two medical applications than adaptation of GPMIL to scalable inference and various scalable MIL algorithms. It also proves to be very competitive in object classification against state-of-the-art adaptations of deep learning to weakly supervised learning.

## 1 Introduction

Improvements in data acquisition technologies make ever larger data masses available for machine learning. However, data annotation capabilities do not keep pace with the amount of available raw data. Recognition of patterns of interest requires ground-truth labels, gathering of which comes at the cost of manpower in most applications. Even worse, annotation can only be done by domain experts in some fields, such as medicine. Weakly supervised learning (WSL) emerged as a remedy to this problem by allowing labels to be provided only for data groups, called *bags*, rather than for individual instances.

Multiple instance learning (MIL) [20] is the most frequent form of WSL for classification. MIL assumes that the bag label is given by the maximum of the labels of the instances

within that bag. That is, a bag with one or more positive instances is deemed positive. Having shown great success in a diverse set of applications, MIL is known to suffer from high sensitivity to false positives. This is because the MIL assumption necessitates *all* instances within a negative bag to be correctly predicted. In addition to studies targeting this specific drawback of MIL [9, 17], there also emerged WSL-flavored approaches to MIL that learn directly to explain the bag label by-passing the intermediary step of discovering the instance labels. A seminal example is mi-Graph [36] which constructs a topological graph for each training bag and learns a bag-level predictor defined on pairwise graph similarities. Soft-Bag SVM [17] is a more recent method that characterizes bags by a weighted linear sum of the instance feature vectors and learns a mapping from the bag-level feature vectors to bag labels. While both of these methods are impressively effective on up to middle-sized data sets, their scalability is severely limited by their either training or prediction time complexities.

We hereby formulate for the first time a weakly supervised Gaussian process that is designed specifically for large-scale prediction tasks. Our model assigns each instance a share (i.e. local evidence) in determination of the bag label. It assumes that these instance-level evidences follow a sparse Gaussian process and their accumulation determines the bag label. We make this unorthodox choice for standard MIL setups since:

- Accumulating instance-level evidence allows learning only to predict bag labels without learning to predict instance labels. Hence it solves an easier problem than MIL, preserving all the expressive power to the target task.

- When applied to Bayesian models such as GPs, the MIL setup necessitates prediction of the bag label from the *maximum* instance evidence and the $\max(\cdot)$ operator does not render effective inference schemes applicable.

- Contrary to the point above, instance-level evidence accumulation (i.e. the $\sum(\cdot)$ operator) enables variational inference consisting only of closed-form updates, which boosts up learning speed and eliminates the need for tuning a proper learning rate.

We build our solution to WSL on GPs due to their multiple advantages for the context. The most important is that, while being kernelized learners with a high degree of non-linearity, they have been shown to scale easily up to millions of data points [10]. A GP-based model can also trivially be adapted to active learning [12] and online learning [6], which are two other machine learning setups to facilitate learning from big data sets.

We compare our variational weakly supervised GP to a large spectrum of alternatives, such as the sparsified extension of the unscalable GPMIL [15], various other scalable MIL algorithms, weakly supervised object detection pipelines, very deep neural nets, and a combination deep learning and MIL. We observe our model to be very competitive on an exhaustively-studied object classification benchmark against many pipelines dedicated for this very task. We also report it to greatly improve the state-of-the-art on two challenging medical image analysis tasks where such well-settled preprocessing pipelines (e.g. pre-trained neural nets) do not exist: i) diabetic retinopathy screening from eye fundus images, and ii) tumor detection from histopathological sentinel lymph node slide images. Lastly, we quantify that our variational inference scheme with closed-form update rules provides nearly as fast convergence as stochastic variational inference (SVI) [11] with a *manually*-tuned learning rate.

### 1.1 Notation

The WSL setup assumes a data set $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ consisting of $N$ instances represented by feature vectors $\mathbf{x}_n$ partitioned into bags as $\mathbf{X} = \{\mathbf{X}_1 \cup \mathbf{X}_1 \cup \cdots \cup \mathbf{X}_B\}$ and the corresponding bag labels $\mathbf{T} = \{T_1, \cdots, T_B\}$ with $T_b \in \{0,1\}$. Our concern is to learn a predictor that takes the observations within a bag as input and predicts the bag label as output. Let $\mathbf{f}$ be an $N$-dimensional vector of latent variables assigned to instances. Then, $\mathbf{f}_b = [f_1^b, \cdots, f_{N_b}^b]$ denotes the $N_b$-dimensional subset of this vector containing its entries corresponding to bag $b$. We denote a multivariate normal distribution by $\mathcal{N}(\mathbf{x}|\mu,\Sigma)$, and a vector of 1's by $\mathbf{1}$, the size of which is determined by the context. Additionally, $\mathbb{E}_{q(\mathbf{x})}[\cdot]$ is the expectation of the argument with respect to distribution $q(\mathbf{x})$, $\mathbb{KL}(\cdot||\cdot)$ is the Kullback-Leibler divergence between the two distributions in its arguments, $[\mathbf{K}_{\mathbf{AB}}]_{ij} = k(\mathbf{a}_i, \mathbf{b}_j|\theta)$ is the Gram matrix between two data sets $\mathbf{A}$ and $\mathbf{B}$ with respect to a kernel function $k$ parameterized by $\theta$, and $diag(\mathbf{A})$ returns a matrix containing the same diagonal elements as the square matrix in its argument and zero elsewhere.

## 2 Scalability of the Prior Art

MI-SVM [2] and mi-SVM [2] are two well-known adaptations of the support vector machine (SVM) to the MIL setting. They are special cases of the latent structured SVM [33] with binary latent nodes assigned to instances from positive bags. For MI-SVM, a positive value appends the instance into the active data set SVM operates on, and a negative value takes it out. MI-SVM allows a single data point per bag to be active at a time. For mi-SVM, the nodes are latent labels of the instances. mi-SVM allows any configuration except all nodes of a positive bag being negative. In both models, unary potentials are a function of the decision margins of the standard SVM on instances. These models also have an $N_b$-ary potential for the latent nodes per bag that assigns infinite cost to the forbidden configurations. Subsequent adaptations of the latent SVM to MIL also exist [8]. Latent SVM, in its native form, poses an NP-Hard problem, to which Yu et al. [33] have found an approximate solution via Concave-Convex Procedure (CCCP) [34]. PC-SVM [2] regularizes false positives by forcing the positive and negative instances to be on different lower-dimensional manifolds. The Soft-Bag SVM [17] represents each bag by a feature vector, which is a linear combination of the feature vectors of the instances within that bag. This combination allows a number of positive instances in negative bags lower than a predetermined threshold. Training of PC-SVM and Soft-Bag SVM involves a quadratically constrained quadratic program (QCQP), which scales polynomially with the data set size [4]. As all the models above are parametric, once trained, their prediction time is independent of the training set size.

Another class of MIL models is non-parametric, and needs to process the entire training data for every prediction. Citation k-NN [29] calculates the Hausdorff distance between the query bag and all the bags in the training set and assigns the bag to the class of the nearest neighbor. The predictive mean of GPMIL [15] involves calculation of the distance of all instances in the query bag to all instances in the training set with respect to a kernel function. Counter-intuitively, mi-Graph [36] is also non-parametric even though it builds on a large-margin solution based on bag-level similarities. The bag-level similarity metric used by mi-Graph is based on pairwise similarities of *all* instances between two bags. Hence, similarities of all instances in a new query bag to all instances in the support vector bags have to be calculated. Finally, scalability of ensemble learning based MIL approaches [32, 35] is

bounded by the chosen weak learner.

The first study on scalable MIL [24] adapts locally sensitive hashing to MIL. Bergeron et al. [4] scale up an existing multi-instance ranking model (MIRank) with a bundle method for non-convex optimization. Li et al. [18] extends the scalability of a weakly supervised large-margin learner via convex relaxation of the mixed-integer programming based optimization of its original loss. A recent work by Wei et al. [30] introduces a scalable MIL algorithm that first clusters input data into groups, then describes the bags by their Fisher information, and finally feeds these bag descriptors into a standard learner.

The recent trend is to approach weakly supervised learning as an application of deep learning. Here, the challenge is to devise the most appropriate deep neural net architecture for the application at hand. Some methods in this line allocate part of their neural net for proposal generation [31]. These proposals are learned jointly with the parameters of upper and lower layers of a large neural net. Some other methods choose to learn progressively more abstract representations of the raw input without discarding its certain regions using proposals [14, 19].

## 3 Variational Weakly Supervised Gaussian Processes

Given a bag $b$ with input $\mathbf{X}_b = [\mathbf{x}_1^b; \cdots; \mathbf{x}_{N_b}^b]$, we assume the contribution of the instances to bag prediction to follow a Gaussian process: $f_n^b \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ with mean function $\mu(\mathbf{x})$ set to zero to keep the prior uninformative and the covariance function $k(\mathbf{x}, \mathbf{x}')$. The binary bag label[1] is then determined after accumulating of the contributions of all its instances: $T_b | f_1^b, \cdots, f_{N_b}^b \sim \text{Bernoulli}(T_b | \sigma(f_1^b + \cdots + f_{N_b}^b))$, where $\sigma(\cdot)$ is a sigmoid function.

As the scalability of the standard GP is limited by its $N \times N$ covariance matrix, we build on its *Fully Independent Training Conditional (FITC)* approximation [27], which decomposes the full kernel matrix into a combination of low-rank and diagonal matrices. This is achieved by the assumption that the entire data set is generated from a small set of pseudo data points, called *inducing points*. A FITC-approximated sparse GP prior on a random vector $\mathbf{f}$ is defined as

$$\begin{aligned} \mathcal{SGP}(\mathbf{f}, \mathbf{u} | \mathbf{X}, \mathbf{Z}) &= p(\mathbf{u}|\mathbf{Z}) p(\mathbf{f}|\mathbf{u}, \mathbf{X}, \mathbf{Z}) \\ &= \mathcal{N}(\mathbf{u}|\mathbf{0}, \mathbf{K_{ZZ}}) \mathcal{N}(\mathbf{f}|\mathbf{K_{XZ}}\mathbf{K_{ZZ}}^{-1}\mathbf{u}, diag(\mathbf{K_{XX}} - \mathbf{K_{XZ}}\mathbf{K_{ZZ}}^{-1}\mathbf{K_{ZX}})), \end{aligned}$$

where $\mathbf{Z} = [\mathbf{z}_1; \cdots; \mathbf{z}_P]$ with $P \ll N$ contains the inducing points in its rows, and $\mathbf{u}$ is called the *inducing output* vector, as its entries correspond to the outputs of the inducing points. FITC approximation converts the non-parametric full GP into a parametric model, which can predict query points based only on the posterior distribution of $\mathbf{u}$, the size of which does not grow with training data size.

Putting the pieces together, we propose

$$\begin{aligned} p(\mathbf{f}, \mathbf{u} | \mathbf{X}, \mathbf{Z}) &= \mathcal{SGP}(\mathbf{f}, \mathbf{u} | \mathbf{X}, \mathbf{Z}), \\ p(\mathbf{T}|\mathbf{f}) &= \prod_{b=1}^{B} \text{Bernoulli}(T_b | \sigma(\mathbf{f}_b^T \mathbf{1})). \end{aligned}$$

The vector $\mathbf{f}_b$ has the instance-level contributions within the bag, which follow a sparse GP prior, and $\sigma(s) = 1/(1+e^{-s})$ is the logistic sigmoid function. These instance-level

[1] Extension to continuous output, hence multiple instance regression, is trivial: $Y_b \sim \mathcal{N}(Y_b | f_1^b + \cdots + f_{N_b}^b, \sigma^2)$.

contributions are accumulated in the term $\mathbf{f}_b^T\mathbf{1}$. As this likelihood does not follow the standard MIL assumption that the maximum instance label determines the bag label, the instance-level contributions do not correspond to instance labels. Rather we devise this model to learn directly to predict bag labels and tailor its formulation specifically for large-scale learning tasks. We call this model the *Variational Weakly Supervised Gaussian Process (VWSGP)*.

The sum operator in the likelihood makes VWSGP amenable for efficient variational inference. Contrary to the early work on scalable GPs [10], we favor closed-form updates over SVI to leverage scalability. While stochastic updates are known to achieve a steeper learning rate in initial iterations, they reach at the optimal solution later than deterministic updates [5]. More critically, in case of variational inference on GPs, stochastic updates can only be done when the forms of the variational factor distributions are pre-imposed and a proper learning rate is used. On the other hand, some closed-form solutions, although they can be made only in batch mode, do not require any tuning on the learning rate. Here we promote the latter and derive a closed-form update rule for our WSL model above, since it is hardly feasible to tune learning rates on big data.

We approximate the intractable Bernoulli likelihood with the Jaakkola bound [13], which is frequent in Bayesian logistic regression, but surprisingly not yet applied to sparse GPs

$$p(T_b|\mathbf{f}) \geq e^{\mathbf{f}_b^T\mathbf{1}T_b}\sigma(\xi_b)e^{-(\mathbf{f}_b^T\mathbf{1}+\xi_b)/2-\lambda(\xi_b)(\mathbf{f}_b^T\mathbf{1}\mathbf{1}^T\mathbf{f}_b-\xi_b^2)},$$

where $\xi_b$ is a new variational parameter assigned to bag $b$ and $\lambda(\xi_b)=\frac{1}{2\xi_b}(\sigma(\xi_b)-1/2)$.

In variational inference, the goal is to approximate the intractable posterior $p(\mathbf{f},\mathbf{u}|\mathbf{T},\mathbf{X})$ with another distribution $Q$, called the variational distribution. We adopt the sparse GP factorization used earlier by Titsias et al. [28] to manifold learning and apply it for the first time to weakly supervised learning: $Q=p(\mathbf{f}|\mathbf{u},\mathbf{Z},\mathbf{X})q(\mathbf{u})$, yet, we do not assume any form for $q(\mathbf{u})$. It is possible to decompose the marginal likelihood into the following three terms:

$$\log p(\mathbf{T}|\mathbf{X}) = \underbrace{\mathbb{E}_Q[\log p(\mathbf{T}|\mathbf{f})]-\mathbb{KL}(q(\mathbf{u})||p(\mathbf{u}))}_{\text{Evidence Lower Bound }(\mathcal{L})}+\mathbb{KL}(Q||p(\mathbf{f},\mathbf{u}|\mathbf{X},\mathbf{T})).$$

In learning, we aim to find the optimal $q(\mathbf{u})$ which minimizes the third term above. Note that this term converges to zero as $Q$ approaches to the true posterior. Since the log-marginal likelihood on the l.h.s. is constant, minimizing the third term is equivalent to maximizing the first two terms, which are together called the *Evidence Lower Bound (ELBO)*, denoted by $\mathcal{L}$. Note that $\mathcal{L}$ is a *functional* to be optimized with respect to the function $q(\mathbf{u})$. Evaluating its gradient on the space of all possible $q(\mathbf{u})$'s at zero gives:

$$\frac{\partial\mathcal{L}}{\partial q(\mathbf{u})}=\mathbb{E}_{p(\mathbf{f}|\mathbf{u},\mathbf{Z},\mathbf{X})}[\log p(\mathbf{T}|\mathbf{f})]-\log q(\mathbf{u})-1=0,$$

leading to the update rule $\log q(\mathbf{u})\leftarrow\mathbb{E}_{p(\mathbf{f}|\mathbf{u},\mathbf{Z},\mathbf{X})}[\log p(\mathbf{T}|\mathbf{f})]$. Evaluating the r.h.s. gives $q(\mathbf{u})\leftarrow\mathcal{N}(\mathbf{u}|\mathbf{m},\mathbf{S})$ with

$$\mathbf{S}\leftarrow\left[\left(\sum_{b=1}^{B}2\lambda(\xi_b)\mathbf{K}_{\mathbf{ZZ}}^{-1}\mathbf{K}_{\mathbf{ZX}_b}\mathbf{1}\mathbf{1}^T\mathbf{K}_{\mathbf{ZX}_b}^T\mathbf{K}_{\mathbf{ZZ}}^{-1}\right)+\mathbf{K}_{\mathbf{ZZ}}^{-1}\right]^{-1},$$

$$\mathbf{m}\leftarrow\mathbf{S}\Big[\sum_{b=1}^{B}(T_b-1/2)\mathbf{K}_{\mathbf{ZZ}}^{-1}\mathbf{K}_{\mathbf{ZX}_b}\mathbf{1}\Big].$$

Table 1: Data sets used in our experiments. Note that VOC 2007 consists of 20 different binary classification problems. Hence, we report the *average* positive and negative bag counts across 20 problems for this data set.

| Task | Instances | Features | +/− Bags |
|---|---|---|---|
| Object Classification (VOC 2007) | 19 923 912 | 100 | 716.3/9246.6 |
| Diabetic Retinopathy Screening | 361 201 | 462 | 83/83 |
| Metastasis Screening in Tissue Slides | 1 004 297 | 657 | 78/90 |

Here, $\mathbf{S}$ is a $P \times P$ matrix, $P$ being the number of inducing points, which is chosen to be very small (e.g. 250). Hence, inversion of $\mathbf{S}$ at every update does not set a bottleneck on training time. We update $\xi_b$'s by evaluating the derivative of the complete ELBO with respect to $\xi_b$ at zero, which yields

$$\xi_b^2 \leftarrow tr\Big(\mathbf{1}\mathbf{1}^T \mathbf{K}_{\mathbf{Z}\mathbf{X}_b}^T \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-1} (\mathbf{m}\mathbf{m}^T + \mathbf{S}) \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-1} \mathbf{K}_{\mathbf{Z}\mathbf{X}_b}\Big) + tr\Big(\mathbf{1}\mathbf{1}^T diag(\mathbf{K}_{\mathbf{X}_b\mathbf{X}_b} - \mathbf{K}_{\mathbf{Z}\mathbf{X}_b}^T \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-1} \mathbf{K}_{\mathbf{Z}\mathbf{X}_b})\Big).$$

The predictive distribution of VWSGP for a new input bag $\mathbf{X}^*$ is

$$p(T^*|\mathbf{X}^*, \mathbf{X}, \mathbf{T}, \mathbf{Z}) = \int\int p(T^*|\mathbf{f}^*) p(\mathbf{f}^*|\mathbf{u}, \mathbf{Z}, \mathbf{X}^*) p(\mathbf{u}|\mathbf{X}, \mathbf{T}) d\mathbf{u} d\mathbf{f}^*.$$

Replacing the intractable posterior $p(\mathbf{u}|\mathbf{X}, \mathbf{T})$ with its variational approximation $q(\mathbf{u}|\mathbf{m}, \mathbf{S})$ inferred during training and taking the inner integral on $\mathbf{u}$ gives

$$p(\mathbf{f}^*|\mathbf{Z}, \mathbf{X}^*) = \mathcal{N}(\mathbf{f}^*|\mathbf{A}^*\mathbf{m}, diag(\Lambda^* + \mathbf{A}^*\mathbf{S}\mathbf{A}^{*T})),$$

where $\mathbf{A}^* = \mathbf{K}_{\mathbf{Z}\mathbf{X}^*}{}^T \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-1}$ and $\Lambda^* = diag(\mathbf{K}_{\mathbf{X}^*\mathbf{X}^*} - \mathbf{K}_{\mathbf{Z}\mathbf{X}^*}^T \mathbf{K}_{\mathbf{Z}\mathbf{Z}}^{-1} \mathbf{K}_{\mathbf{Z}\mathbf{X}^*})$. We approximate the intractable outer integral on $\mathbf{f}^*$ by Monte Carlo integration

$$\mathbf{f}_{(i)}^* \sim p(\mathbf{f}^*|\mathbf{Z}, \mathbf{X}^*), \quad i = 1, \cdots, S$$

$$p(T^*|\mathbf{X}^*, \mathbf{X}, \mathbf{T}, \mathbf{Z}) \approx \frac{1}{S} \sum_{i=1}^{S} p(T^*|\mathbf{f}_{(i)}^*).$$

Since $p(\mathbf{f}^*|\mathbf{Z}, \mathbf{X}^*)$ has diagonal covariance, it is possible to sample independently for each instance $f_b^n$ from a univariate normal distribution. Since the samples are taken from the true distribution, a small sample set (e.g. 100) suffices for stable predictions [23]. This predictor corresponds to a voting Bayes classifier, which is proven to tightly approximate the ideal Bayes classifier [21].

# 4 Results

We evaluate VWSGP on three hard computer vision tasks: i) object classification from natural images, ii) diabetic retinopathy screening, and iii) detection of potentially tiny metastatic tumors on sentinel lymph node tissues from histopathology whole slide images. Numeric details of the used data sets are given in Table 1.

Table 2: Average Precision (AP) scores for object classification on VOC 2007 test split.

| Object | VWSGP | VGG-S [14] | DMIL [31] | CNN-SVM[22] | 1000C [19] |
|---|---|---|---|---|---|
| aeroplane | 93.3 | **95.3** | 93.5 | 91.2 | 88.5 |
| bicycle | 89.9 | **90.4** | 83.4 | 81.4 | 81.5 |
| bird | 88.9 | **92.5** | 86.9 | 82.1 | 87.9 |
| boat | 86.8 | **89.6** | 83.6 | 81.1 | 82.0 |
| bottle | **67.9** | 54.4 | 54.2 | 51.6 | 47.5 |
| bus | **86.7** | 81.9 | 81.6 | 81.6 | 75.5 |
| car | **91.5** | **91.5** | 86.6 | 84.4 | 90.1 |
| cat | 90.7 | **91.9** | 85.2 | 83.9 | 87.2 |
| chair | **64.8** | 64.1 | 54.5 | 54.5 | 61.6 |
| cow | **83.8** | 76.3 | 68.9 | 61.0 | 75.7 |
| dining table | **78.5** | 74.9 | 53.8 | 53.8 | 67.3 |
| dog | 81.8 | **89.7** | 73.2 | 72.3 | 85.5 |
| horse | **93.2** | 92.2 | 78.8 | 74.9 | 83.5 |
| motorbike | 86.3 | **86.9** | 79.0 | 75.6 | 80.0 |
| person | 93.8 | **95.2** | 86.6 | 83.7 | 95.6 |
| potted plant | **68.3** | 60.7 | 51.2 | 47.4 | 60.8 |
| sheep | 82.5 | **82.9** | 74.4 | 71.7 | 76.8 |
| sofa | **71.1** | 68.0 | 63.7 | 60.0 | 58.0 |
| train | 92.7 | **95.5** | 91.5 | 88.3 | 90.4 |
| tv monitor | **81.7** | 74.4 | 80.4 | 79.4 | 77.9 |
| **mAP** | **83.7** | 82.4 | 75.5 | 73.0 | 77.7 |

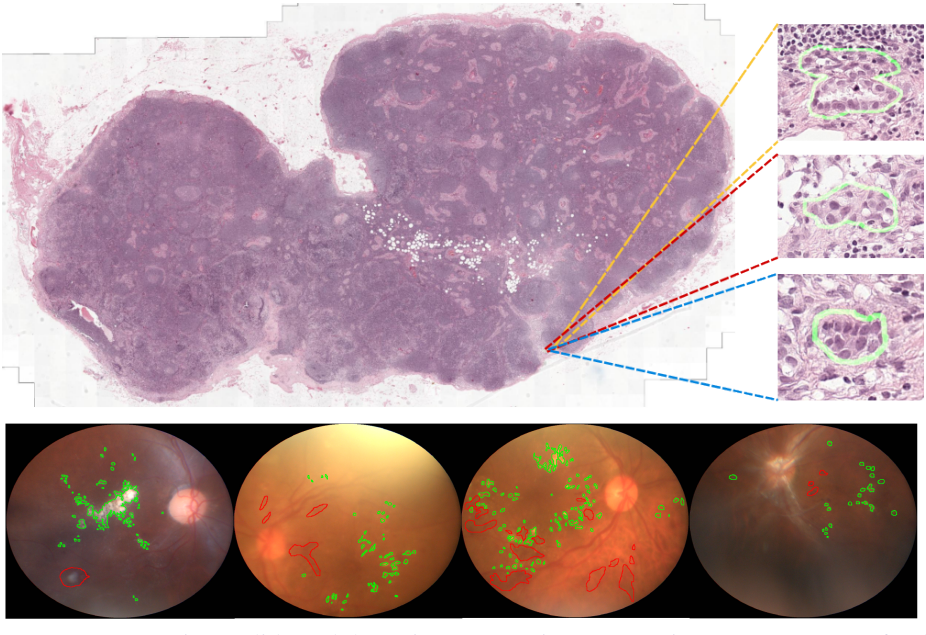

Figure 1: **Top:** A tissue slide and three tiny metastatic tumor regions. **Bottom:** Eye fundus images, where diabetic retinopathy lesions are marked (red: hemorrhages, green: exudates).

### 4.1 Object classification on VOC 2007

In Table 2, we compare our VWSGP to four state-of-the-art mainstream object classification pipelines on the VOC 2007 benchmark data set. All these four models train a deep neural net either with or without the notion of region proposal. In any case, all parameters of these models are learned in a unified optimization scheme. Similarly to these models, we use a neural net [26] trained on the external MS COCO data set for proposal generation. However, differently from all these models, we train a non-deep Bayesian kernelized learner on top of the proposal as a disjoint consecutive step. We treat each image as a bag and up to highest-ranking 2000 proposals as its instances. We use the activations of the highest fully-connected layer behind the proposal outputs as instance features. We reduce the feature dimensionality to 100 using principal component analysis and use RBF as the kernel function with length scale 20. Our generic weakly supervised learner performs very competitively compared to these pipelines tuned tightly to this very application. We keep speed benchmarking out of our scope and satisfy with reporting that VWSGP performs 10 closed-form updates on the VOC 2007 training split (10 020 341 instances) in 162 minutes and predicts on its test split (9 903 571 instances) in 27 minutes. Observing the sparse variant of GPMIL [15] to perform far worse than models above, we also omit a comparison to it for brevity.

### 4.2 Metastasis and diabetes screening

Table 3: Bag-level prediction accuracies (% Acc.) and average precision scores (AP) of models in comparisons on the two medical data sets.

| | Diabetes | | Metastasis | |
|---|---|---|---|---|
| **Model** | **% Acc.** | **AP** | **% Acc.** | **AP** |
| **VWSGP (Ours)** | **91.3** | **0.98** | **61.7** | **0.68** |
| **GPMIL-S** [15] | 80.8 | 0.86 | 53.7 | 0.61 |
| **e-MIL** [16] | 76.0 | 0.93 | 59.8 | 0.61 |
| **mi-FV** [30] | 88.3 | 0.92 | 54.1 | 0.48 |
| **BMIL** [25] | 73.7 | 0.79 | 57.9 | 0.58 |

**Metastasis screening from histopathology tissue images.** Here the task is to classify whole-slide histopathology images as metastatic and healthy. Those images are typically very big and contain the object of interest (i.e metastatic tumor) only in extremely small regions (see Figure 1 (top)), making it a tedious task for pathologists to point them out. It is also commonplace to gather multiple samples from one patient and multiple images from each of these samples, one per slice. A tool that can predict which slide to start searching for the disease would save the precious time of the pathologist. We treat each $216\,000 \times 95\,304$-pixel whole-slide image as a bag and each $500 \times 500$-pixel patch on its foreground as an instance. We represent each color channel of each patch by a 26-bin intensity histogram, mean of SIFT descriptors extracted from the keypoints within the patch, a 58-dimensional vector of LBP features from 50-pixel cells, and box counts for grid sizes 2 to 8.

**Diabetic retinopathy screening from fundus images.** Diabetes has known effects on the eye, called diabetic retinopathy (DR). With the technology of eye fundus imaging, both its existence can be detected from the eye, and its damage on the eye can be diagnosed. These

detections can be automatized using a system that classifies the fundus image as diseased or healthy, enabling to cheap screening of many individuals, and bringing the positive cases to the attention of an opthalmologist. This could save the eyesight of many people especially in countries with insufficient number of opthalmologists. Typical DR lesions are microaneurysms (small blood spots), hemorrhages (larger bleeding regions), and hard exudates (bright yellow spots). Four sample images with DR lesions are shown in Figure 1 (bottom). We treat each $1920 \times 1440$-pixel eye fundus image as a bag, and each SLIC [1] superpixel on the foreground as an instance. We represent each color channel of each superpixel by a 26-bin intensity histogram and the mean of the SIFT descriptors within that superpixel.

The standard weakly supervised learning methods in the computer vision literature are tailored for natural images. These methods necessitate either a pre-trained neural net on a similar problem [26] or an application-specific region proposal generator [7]. In absence of both in the two medical image analysis problems of our concern, we compare VWSGP against three scalable MIL methods:

- **GPMIL-S:** Our adaptation of Kim et al.'s work [15] to scalable learning (Laplace approximation on a sparse Gaussian process). We mean hereby to test how the accumulated evidence assumption compares to MIL.
- **mi-FV [30]:** One of the most recent MIL algorithms devised specifically for big data.
- **e-MIL [16]:** A state-of-the-art MIL algorithm that performs bag modeling and bag-level classification as two disjoint steps.
- **BMIL [25]:** Adaptation of logistic regression to MIL, which finds a linear decision boundary between bag classes. We choose this baseline for two reasons: i) to show how well an earlier application of Bayesian MIL on computer-assisted diagnosis generalizes to our medical data sets, and ii) to illustrate how the non-linear decision boundaries provided by the GP prior boosts up prediction performance.

We perform 20 replications of a randomly chosen 50%-50% train/test split and report the AP scores as well as plain accuracies averaged across these 20 trials in Table 3. For VWSGP and GPMIL-S, we use 50 inducing points for the diabetic retinopathy data set and 250 for the larger metastasis data sets, and use the radial basis function (RBF) as the kernel function. We set the kernel length scale to the square-root of the feature dimensionality. Inducing points are fixed to k-means cluster centroids as in [10]. We train VWSGP with 50 closed-form updates on the entire training set and GPMIL-S on $50 \times B_t$ stochastic updates, $B_t$ being the number of training bags. We used the source code provided by the authors for e-MIL and mi-FV, and implemented BMIL from the details given in the paper.

## 4.3 Closed-Form Updates versus SVI

Here we compare the convergence rate of our closed-form update based learning scheme with stochastic variational inference (SVI). For the latter, we adapt our VWSGP model to SVI by adding white noise on top of the bag evidence $p(Y_b|\mathbf{f}_b) = \mathcal{N}(Y_b|\mathbf{f}_b^T\mathbf{1}, \alpha^{-1})$ and using a probit link function $\Phi(s) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{s} e^{-\frac{1}{2}v^2} dv$ in the likelihood $p(T_b|Y_b) = \text{Bernoulli}(T_b|\Phi(Y_b))$ in place of the logistic sigmoid above. Probit enables tractable integration of $Y_b$, hence, allows mapping real-valued output to binary. The posterior is then approximated by a variational distribution similarly to above $Q = p(\mathbf{f}|\mathbf{u},\mathbf{Z},\mathbf{X})q(\mathbf{u})$ where $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}|\mathbf{m},\mathbf{S})$ is this time

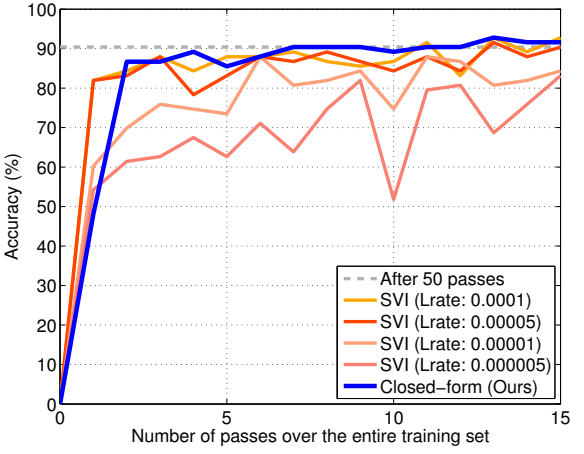


Figure 2: Our closed-form update rule follows a very similar learning curve to the one for SVI with a *suitable* learning rate, which is discovered only after trial-and-error.

imposed to be normal distributed. The ELBO is maximized by gradient ascent with respect to $\mathbf{m}$ and $\mathbf{S}$. Figure 2 shows the learning curves of the closed-form update based model we propose (blue) and SVI-based inference with different learning rates on one train/test split of the diabetic retinopathy data set. Firstly, our model converges to the sink point after only six passes over the training data, which ensures scalability of our model to big data. Lastly, its learning curve tightly approximates SVI-based inference with a *suitable* learning rate, which can be discovered only after several trials. Our approach clears away the obligation to search for a suitable learning rate, which is obviously prohibitive when analyzing large data volumes.

# 5 Discussion

The fact that VWSGP clearly outperforms GPMIL-S serves as a proof-of-concept for the effectiveness of the accumulated evidence approach. This outcome is also consistent with the Vapnik's razor principle: *"When solving a (learning) problem of interest, do not solve a more complex problem as an intermediate step"*. Seeing both sides of the coin, VWSGP is clearly more effective than GPMIL-S in bag label prediction, however it is not capable of predicting instance labels while GPMIL-S is (although never benchmarked till present).

Competitiveness of VWSGP in object classification when applied on top of pre-trained region proposal generators praises weakly supervised learning research as an area to be also explored independently from deep neural nets. With a good weakly supervised predictor, it is indeed possible to challenge deep learners trained on cropped image regions coming from a proposal generator [22, 31] or on the entire input [14, 19].

As for the medical applications, improvement of VWSGP over e-MIL and mi-FV can be attributed to that VWSGP learns how to infer instance-level evidence and how to predict the bag label from this evidence *jointly*. VWSGP outperforms BMIL with a large margin thanks to the kernelized GP prior, which enables learning non-linear decision boundaries.

# References


[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002.

[3] C. Bergeron, G. Moore, J. Zaretzki, C.M. Breneman, and K.P. Bennett. Fast bundle algorithm for multiple-instance learning. *Pattern Analysis and Machine Intelligence*, 34(6):1068–1079, 2012.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] M.P. Friedlander and M. Schmidt. Hybrid deterministic-stochastic methods for data fitting. *SIAM Journal on Scientific Computing*, 34(3):A1380–A1405, 2012.

[6] R. Frigola, Y. Chen, and C. Rasmussen. Variational Gaussian process state-space models. In *NIPS*, 2014.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[8] H. Hajimirsadeghi, J. Li, G. Mori, M. Zaki, and T. Sayed. Multiple instance learning by discriminative training of Markov networks. In *UAI*, 2013.

[9] Y. Han, Q. Tao, and J. Wang. Avoiding false positive in multi-instance learning. In *NIPS*, 2010.

[10] J. Hensman, N. Fusi, and N.D. Lawrence. Gaussian processes for big data. In *UAI*, 2013.

[11] M.D. Hoffman, D.M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[12] N. Houlsby, F. Huszar, Z. Ghahramani, and J.M. Hernández-Lobato. Collaborative Gaussian processes for preference learning. In *NIPS*, 2012.

[13] T.S. Jaakkola and M.I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

[14] A. Vedaldi A. Zisserman K. Chatfield, K. Simonyan. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.

[15] M. Kim and F. de la Torre. Gaussian processes multiple instance learning. In *ICML*, 2010.

[16] G. Krummenacher, C.S. Ong, and J. Buhmann. Ellipsoidal multiple instance learning. In *ICML*, 2013.

[17] W. Li and N. Vasconcelos. Multi-instance learning for soft bags via top instances. In *CVPR*, 2015.

[18] Y.-F. Li, I.W. Tsang, J.T. Kwok, and Z.-H. Zhou. Convex and scalable weakly labeled SVMs. *Journal of Machine Learning Research*, 14(1):2151–2188, 2013.

[19] I. Laptev J. Sivic M. Oquab, L. Bottou. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.

[20] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. *NIPS*, 1998.

[21] A.Y. Ng and M.I. Jordan. Convergence rates of the voting Gibbs classifier, with application to bayesian feature selection. In *ICML*, 2001.

[22] X. Zhang M. Mathieu R. Fergus Y. LeCun P. Sermanet, D. Eigen. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014.

[23] J. Paisley, D. Blei, and M. Jordan. Variational Bayesian inference with stochastic search. In *ICML*, 2012.

[24] W. Ping, Y. Xu, J. Wang, and X.-S. Hua. FAMER: making multi-instance learning better and faster. In *SDM*, 2011.

[25] V.C. Raykar, B. Krishnapuram, J. Bi, M. Dündar, and R.B. Rao. Bayesian multiple instance learning: Automatic feature selection and inductive transfer. In *ICML*, 2008.

[26] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[27] E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *NIPS*, 2006.

[28] M.K. Titsias and N.D. Lawrence. Bayesian Gaussian process latent variable model. In *AISTATS*, 2010.

[29] J. Wang and J.-D. Zucker. Solving multiple-instance problem: A lazy learning approach. In *ICML*, 2000.

[30] X.-S. Wei, J. Wu, and Z.-H. Zhou. Scalable multi-instance learning. In *ICDM*, 2014.

[31] J. Wu, Y. Yinan, C. Huang, and Y. Kai. Deep multiple instance learning for image classification and auto-annotation. In *CVPR*, 2015.

[32] X. Xu and E. Frank. Logistic regression and boosting for labeled bags of instances. In *Advances in Knowledge Discovery and Data Mining*, pages 272–281. 2004.

[33] C.-N.J. Yu and T. Joachims. Learning structural SVMs with latent variables. In *ICML*, 2009.

[34] A.L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.

[35] C. Zhang, J.C. Platt, and P.A. Viola. Multiple instance boosting for object detection. In *NIPS*, 2005.

[36] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li. Multi-instance learning by treating instances as non-iid samples. In *ICML*, 2009.