Supplementary Material for the Paper: ”Variational Bayesian Multiple Instance Learning with Gaussian Processes”

Manuel Haußmann[1] Fred A. Hamprecht[1] Melih Kandemir[1,2*]

[1]HCI/IWR, Heidelberg University [2]Özyeğin University

{manuel.haussmann, fred.hamprecht}@iwr.uni-heidelberg.de

melih.kandemir@ozyegin.edu.tr

1. Updates for VGPMIL

As we derive the updates of the variational distributions, we drop all independent terms from the equations and use the notation $a \overset{c}{=} b$ to mean $a = b$ up to an additive constant.

Updating $q(u)$**.** Equation 9 gives us as the optimal update for $q(u)$

$$\log q(u) \overset{c}{=} \langle \log p(y|f) \rangle + \log p(u).$$

To derive an analytical solution we need to apply the Jaakkola bound on the first term

$$\begin{aligned}
\langle \log p(y|f) \rangle &= \Big\langle \sum_n \log p(y_n|f_n) \Big\rangle \\
&\geq \sum_n \langle y_n \rangle \langle f_n \rangle - \frac{\langle f_n \rangle + \xi_n}{2} \\
&\quad - \lambda(\xi_n) \left(\langle f_n^2 \rangle - \xi_n^2 \right) \\
&\overset{c}{=} \sum_n \langle f_n \rangle \left(\langle y_n \rangle - \tfrac{1}{2} \right) - \lambda(\xi_n) \langle f_n^2 \rangle \\
&= \langle f \rangle^\top \left(\langle y \rangle - \tfrac{1}{2} \right) - \tfrac{1}{2} \langle f^\top \Lambda f \rangle,
\end{aligned}$$

where $\Lambda = 2\mathrm{diag}\big((\lambda(\xi_1), ..., \lambda(\xi_N))\big)$. The complete expression then becomes[1]

$$\begin{aligned}
\log q(u) &\overset{c}{=} \langle f \rangle^\top \left(\langle y \rangle - \tfrac{1}{2} \right) - \tfrac{1}{2} \langle f^\top \Lambda f \rangle - \tfrac{1}{2} u^\top K_{ZZ}^{-1} u \\
&\overset{c}{=} u^\top K_{ZZ}^{-1} K_{ZX} \left(\langle y \rangle - \tfrac{1}{2} \right) - \tfrac{1}{2} u^\top K_{ZZ}^{-1} u \\
&\quad - \tfrac{1}{2} u^\top K_{ZZ}^{-1} K_{ZX} \Lambda K_{XZ} K_{ZZ}^{-1} u,
\end{aligned}$$

and we get the desired $q(u) = \mathcal{N}(u|m, S)$ with

$$\begin{aligned}
S &\leftarrow \left(K_{ZZ}^{-1} K_{ZX} \Lambda K_{XZ} K_{ZZ}^{-1} + K_{ZZ}^{-1} \right)^{-1}, \\
m &\leftarrow S K_{ZZ}^{-1} K_{ZX} \left(\langle y \rangle - \tfrac{1}{2} \right).
\end{aligned}$$

The optimal update for ξ_n^2 can be derived by taking the derivative of the Jaakkola bound.

The $\max$ **decomposition (Equation 13).** The variational inference updates for $q(y_n)$ in the paper rely on our ability to separate instance y_n from the others in bag b ($\{y_i\}_{b-n}$) in the expression $\max\{y_i\}_b$. To achieve this we introduced the following decomposition for the $\max$ over binary y_i

$$\max\{y_i\}_b = y_n + \max\{y_i\}_{b-n} - y_n \max\{y_i\}_{b-n}.$$

To proof this equation we use the noisy OR function, *i.e.* that for a set of binary values $\{y_1, ..., y_N\} \in \{0, 1\}^N$ we have

$$\max\{y_1, ..., y_N\} = 1 - \prod_{i=1}^{N} (1 - y_i).$$

With this, we get

$$\begin{aligned}
\max\{y_i\}_b &= 1 - \prod_i (1 - y_i) = 1 - (1 - y_n) \prod_{i \neq n} (1 - y_i) \\
&= 1 - \prod_{i \neq n} (1 - y_i) + y_n - y_n + y_n \prod_{i \neq n} (1 - y_i) \\
&= \max\{y_i\}_{b-n} + y_n - y_n \Big(1 - \prod_{i \neq n} (1 - y_i) \Big) \\
&= y_n + \max\{y_i\}_{b-n} - y_n \max\{y_i\}_{b-n}.
\end{aligned}$$

Updating $q(y_n)$**.** Starting from Equation 9 we have, assuming instance n to be in bag b, that

$$\begin{aligned}
\log q(y_n) &\overset{c}{=} \langle \log p(T_b | \{y_i\}_b) + \log p(y_n|f_n) \rangle \\
&\overset{c}{=} \langle G_b \rangle \log H + \langle \log p(y_n|f_n) \rangle.
\end{aligned}$$

The second term evaluates to

$$\langle \log p(y_n|f_n) \rangle = y_n \langle f_n \rangle + \text{const}, \qquad (*)$$

[*] The main part of this work has been done while the author was with Heidelberg Collaboratory for Image Processing (HCI), Heidelberg University

[1] Where we use the trace trick to evaluate the expectation $\langle f^\top \Lambda f \rangle$. That is $\langle f^\top \Lambda f \rangle = \langle \mathrm{tr}(\Lambda f f^\top) \rangle = \mathrm{tr}(\Lambda \langle f f^\top \rangle)$.

while the first term can be extended as

$$\begin{aligned}\log H \cdot \langle G_b \rangle &= \log H \cdot \Big\langle T_b \max\{y_i\}_b \\ &\qquad + (1 - T_b)(1 - \max\{y_i\}_b) \Big\rangle \\ &\overset{c}{=} \log H \cdot \big(2T_b \langle \max\{y_i\}_b \rangle - \langle \max\{y_i\}_b \rangle \big).\end{aligned}$$

In order to separate the instance y_n from the other instances in bag b, we use the $\max$ decomposition from Equation 13. This gives

$$\begin{aligned}\log H \cdot \langle G_b \rangle &\overset{c}{=} \log H \cdot \Big(2T_b \big(y_n + \langle \max\{y_i\}_{b-n} \rangle \\ &\qquad - y_n \langle \max\{y_i\}_{b-n} \rangle \big) - y_n \\ &\qquad - \langle \max\{y_i\}_{b-n} \rangle + y_n \langle \max\{y_i\}_{b-n} \rangle \Big) \\ &\overset{c}{=} \log H \cdot \Big(2T_b \big(y_n - y_n \langle \max\{y_i\}_{b-n} \rangle \big) \\ &\qquad - y_n + y_n \langle \max\{y_i\}_{b-n} \rangle \Big) \\ &\overset{c}{=} y_n \log H \cdot \Big(2T_b - 2T_b \langle \max\{y_i\}_{b-n} \rangle \\ &\qquad + \langle \max\{y_i\}_{b-n} \rangle - 1 \Big). \qquad (**)\end{aligned}$$

We combine $(*)$ and $(**)$ to arrive at the desired update of $q(y_n) = \mathcal{Ber}(y_n|\pi_n)$ with

$$\begin{aligned}\pi_n \leftarrow \sigma\Big(&\langle f_n \rangle + \log H \cdot \big(2T_b + \langle \max\{y_i\}_{b-n} \rangle \\ &- 2T_b \langle \max\{y_i\}_{b-n} \rangle - 1 \big) \Big).\end{aligned}$$

2. Updates for LM-VGPMIL

Our variational distribution for the LM-VGPMIL model is given by $Q = q(u)p(f|u)\prod_n q(y_n)q(g_n)$. However, as mentioned in the main text, to still be able to perform closed form updates, we need to approximate $|f_n|$. We approximate it by $(2y_n - 1)f_n$, where the first factor is the (latent) instance label transformed to $\{\pm 1\}$ to agree with the sign of f_n. Note that this expression poses the problem that we create a circular/loopy structure where y and g both depend on each other, resulting in a directed cyclical graphical model [10]. Unfortunately this means that we lose the guarantees provided by variational inference—a non-decreasing ELBO in each iteration—but our experiments show that the model is still able to learn in practice.

Since we now have a second Bernoulli distribution, we need to apply the bound on the sigmoid twice. We will denote the variational parameters introduced for each instance by the necessary bounds on the Bernoulli distribution in Equation 19 by ξ_n and those for Equation 20 by ϕ_n. Apart from this second bound the updates are derived completely analogously to their non-margin relatives and we will only state the final updates.

Updating $q(u)$. As in our main model Equation 9 gives us the optimal update of $q(u)$ to be

$$\begin{aligned}q(u) &= \mathcal{N}(u|m, S) \qquad \text{with} \\ S &\leftarrow \left(K_{ZZ}^{-1} K_{ZX} \Lambda K_{XZ} K_{ZZ}^{-1} + K_{ZZ}^{-1} \right)^{-1}, \\ m_n &\leftarrow S K_{ZZ}^{-1} K_{Zx_n} \Big(C(2\langle y_n \rangle - 1) \left(\langle g_n \rangle - \tfrac{1}{2} \right) \\ &\qquad + 2C^2 \lambda(\xi_n)(2\langle y_n \rangle - 1)V \\ &\qquad + \langle g_n \rangle \left(\langle y_n \rangle - \tfrac{1}{2} \right) \Big),\end{aligned}$$

where $\Lambda = 2\text{diag}\left((\ldots, \lambda(\xi_n)C^2 + \lambda(\phi_n)\langle g_n \rangle, \ldots) \right)$.

Updating $q(g_n)$ and $q(y_n)$. For $q(g_n)$ we get

$$\begin{aligned}q(g_n) &= \mathcal{Ber}(g_n|\tau_n) \qquad \text{with} \\ \tau_n &\leftarrow \sigma\Big(C\left(\langle (2\langle y_n \rangle - 1) f_n \rangle - V \right) \\ &\qquad + \langle f_n \rangle \left(\langle y_n \rangle - \tfrac{1}{2} \right) - \lambda(\phi_n)\langle f_n^2 \rangle \Big).\end{aligned}$$

And for $q(y_n)$ we have

$$\begin{aligned}q(y_n) &= \mathcal{Ber}(y_n|\pi_n) \qquad \text{with} \\ \pi_n &\leftarrow \sigma\Big(2\langle g_n \rangle C \langle f_n \rangle - C\langle f_n \rangle + 4\lambda(\xi_n)C^2\langle f_n \rangle V \\ &\qquad + \langle f_n \rangle \langle g_n \rangle + \log H \Big(2T_b - 2T_b \langle \max\{y_i\}_{b-n} \rangle \\ &\qquad - 1 + \langle \max\{y_i\}_{b-n} \rangle \Big) \Big),\end{aligned}$$

assuming instance n is in bag b.

Updating ξ_n, ϕ_n. Lastly, the updates for the newly introduced variational parameters are the following:

$$\begin{aligned}\xi_n^2 &\leftarrow C^2 \left(\langle f_n^2 \rangle - 2(2\langle y_n \rangle - 1)\langle f_n \rangle V + V^2 \right) \\ \phi_n^2 &\leftarrow \langle f_n^2 \rangle \langle g_n \rangle.\end{aligned}$$

3. Tables & Figures

Tables 1-5 on the following pages report the class based performances on the PASCAL VOC and Newsgroup data sets. Figure 1 gives t-SNE visualizations, Figure 2 the learning rates for the three smaller data sets and Figure 3, 4 visualize the influence of the additional hyperparameters C, V for the large margin version of the model.

Model	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv	avg
Cinbis *et al.* [2]	67.1	66.1	49.8	34.5	23.3	68.9	**83.5**	44.1	27.7	71.8	**49.0**	48.0	65.2	79.3	37.4	42.9	65.2	51.9	62.8	46.2	54.2
Teh *et al.* [11]	**84.0**	64.6	70.0	**62.4**	25.8	**80.7**	73.9	**71.5**	35.7	81.6	46.5	71.3	79.1	78.8	**56.7**	34.3	69.8	56.7	77.0	**72.7**	64.6
Kantorov *et al.* [4]	83.3	68.6	54.7	23.4	18.3	73.6	74.1	54.1	8.6	65.1	47.1	59.5	67.0	83.5	35.3	39.9	67.0	49.7	63.5	65.2	55.1
Shi and Ferrari [9]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	64.7
Bilen and Vedaldi [1]	68.9	**68.7**	65.2	42.5	40.6	72.6	75.2	53.7	29.7	68.1	33.5	45.6	65.9	**86.1**	27.5	44.9	76.0	**62.4**	66.3	66.8	58.0
Li *et al.* [6]	78.2	67.1	61.8	38.1	36.1	61.8	78.8	55.2	28.5	68.8	18.5	49.2	64.1	73.5	21.4	47.4	64.6	22.3	60.9	52.3	52.4
VGPMIL (ours)	81.5	62.1	68.5	45.3	**68.0**	80.1	58.3	64.7	**54.8**	**84.4**	34.5	71.0	**80.5**	63.7	53.2	**52.7**	**82.3**	**62.4**	**80.1**	71.1	**66.0**
LM-VGPMIL (ours)	78.6	63.0	**71.2**	44.2	65.2	76.3	63.4	63.2	49.0	75.2	17.5	**72.7**	74.6	58.4	52.4	48.6	70.8	55.9	78.9	70.3	62.5

Table 1: PASCAL VOC2007 CorLoc. Results on the trainval set, separated over all classes.

Model	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv	avg
Cinbis *et al.* [2]	39.3	43.0	28.8	20.4	8.0	45.5	47.9	22.1	8.4	33.5	23.6	29.2	38.5	47.9	20.3	20.0	35.8	30.8	41.0	20.1	30.2
Teh *et al.* [11]	48.8	45.9	37.4	**26.9**	9.2	50.7	43.4	43.6	10.6	35.9	27.0	38.6	48.5	43.8	24.7	12.1	29.0	23.2	48.8	41.9	34.5
Kantorov *et al.* [4]	57.1	52.0	31.5	7.6	11.5	55.0	53.1	34.1	1.7	33.1	**49.2**	42.0	47.3	56.6	15.3	12.8	24.8	48.9	44.4	47.8	36.3
Shi and Ferrari [9]	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	37.2
Bilen and Vedaldi [1]	46.4	**58.3**	35.5	25.9	14.0	**66.7**	53.0	39.2	8.9	41.8	26.6	38.6	44.7	59.0	10.8	17.3	40.7	**49.6**	56.9	50.8	39.3
Li *et al.* [6]	54.5	47.4	41.3	20.8	17.7	51.9	**63.5**	**46.1**	21.8	57.1	22.1	34.4	50.5	**61.8**	16.2	**29.9**	40.7	15.9	55.3	40.2	39.5
VGPMIL (ours)	**57.9**	45.8	44.2	23.6	**54.8**	54.7	39.3	40.4	**34.9**	**63.7**	19.0	**53.4**	**61.4**	50.1	**31.1**	27.3	**63.0**	45.0	**59.1**	**52.9**	**46.1**
LM-VGPMIL (ours)	51.9	47.1	**46.4**	17.9	52.3	48.4	41.6	38.8	31.8	59.6	7.4	53.2	53.8	47.5	30.6	28.4	56.2	41.3	56.9	51.9	43.1

Table 2: PASCAL VOC2007 Detection. Results on the test set, separated over all classes.

Model	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv	avg
Kantorov *et al.* [4]	**78.3**	**70.8**	52.5	34.7	36.6	80.0	58.7	38.6	27.7	71.2	**32.3**	48.7	76.2	**77.4**	16.0	48.4	69.9	**47.5**	66.9	**62.9**	54.8
VGPMIL (ours)	59.6	50.9	68.4	**42.7**	**57.2**	75.8	66.3	**64.4**	**47.2**	74.6	19.0	67.8	78.2	66.3	56.7	45.4	75.1	33.3	62.9	54.6	58.3
LM-VGPMIL (ours)	67.9	70.7	**72.7**	35.6	47.6	**81.9**	**72.7**	57.2	36.2	**75.6**	11.3	**69.9**	**84.4**	76.8	**66.0**	**51.0**	**80.3**	30.8	**72.1**	54.8	**60.8**

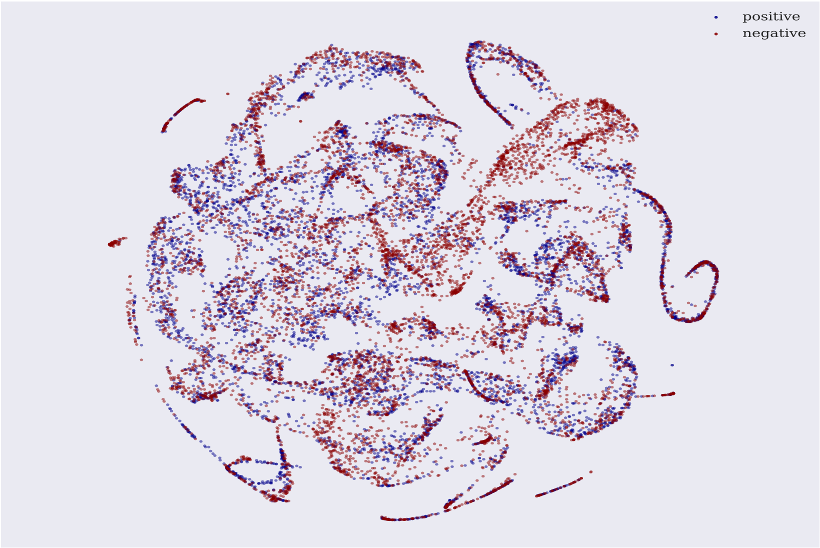
Table 3: PASCAL VOC2012 CorLoc. Results on the trainval set, separated over all classes.

Model	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv	avg
Kantorov *et al.* [4]	**64.0**	**54.9**	36.4	8.1	12.6	53.1	40.5	28.4	6.6	35.3	**34.4**	**49.1**	42.6	**62.4**	19.8	15.2	27.0	**33.1**	33.0	**50.0**	35.3
VGPMIL† (ours)	29.4	30.9	37.2	**15.4**	**35.4**	47.4	42.5	**41.4**	**22.6**	53.3	5.1	43.4	54.7	42.7	35.9	20.1	54.0	17.1	40.0	23.8	34.6
LM-VGPMIL‡ (ours)	39.4	49.7	**42.1**	11.3	13.8	**55.4**	**48.5**	35.7	13.4	**57.3**	2.7	48.3	**63.7**	56.0	**50.5**	**25.2**	**61.1**	13.1	**46.3**	22.8	**37.8**

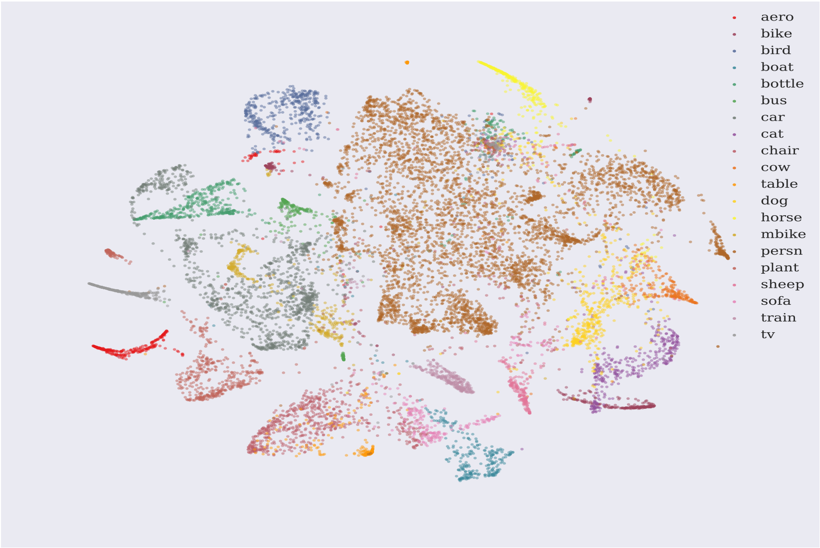
Table 4: PASCAL VOC2012 Detection. Results on the test set, separated over all classes.
†http://host.robots.ox.ac.uk:8080/anonymous/SO7RUO.html
‡http://host.robots.ox.ac.uk:8080/anonymous/9Q2KP6.html

Data Set	LM-VGPMIL (ours)	VGPMIL (ours)	GICF [5]	DPMIL [3]	VF [7]	VFr [7]	GPMIL [3]
alt.atheism	**0.70**	0.67	–	0.67	–	–	0.44
comp.graphics	**0.79**	0.54	–	**0.79**	–	–	0.49
comp.os.ms-windows.misc	**0.52**	0.38	–	0.51	–	–	0.36
comp.sys.ibm.pc.hardware	**0.70**	0.55	–	0.67	–	–	0.35
comp.sys.mac.hardware	**0.79**	0.66	–	0.76	–	–	0.54
comp.windows.x	0.69	0.64	–	**0.73**	–	–	0.36
misc.forsale	**0.54**	0.48	–	0.45	–	–	0.33
rec.autos	0.71	0.46	–	**0.76**	–	–	0.38
rec.motorcycles	**0.76**	0.69	–	0.69	–	–	0.46
rec.sport.baseball	**0.76**	**0.76**	–	0.74	–	–	0.38
rec.sport.hockey	**0.94**	**0.94**	–	0.91	–	–	0.43
sci.crypt	0.82	**0.88**	–	0.68	–	–	0.31
sci.electronics	**0.92**	0.80	–	0.90	–	–	0.71
sci.med	**0.73**	0.65	–	**0.73**	–	–	0.32
sci.space	**0.74**	0.69	–	0.70	–	–	0.32
soc.religion.christian	**0.73**	0.69	–	0.72	–	–	0.45
talk.politics.guns	**0.72**	0.67	–	0.64	–	–	0.38
talk.politics.mideast	**0.87**	0.84	–	0.80	–	–	0.46
talk.politics.misc	**0.64**	0.61	–	0.60	–	–	0.29
talk.religion.misc	0.49	0.45	–	**0.51**	–	–	0.32
Average	**0.73**	0.65	0.71	0.70	0.67	0.59	0.40

Table 5: 20 Newsgroups database. Results for each data set.

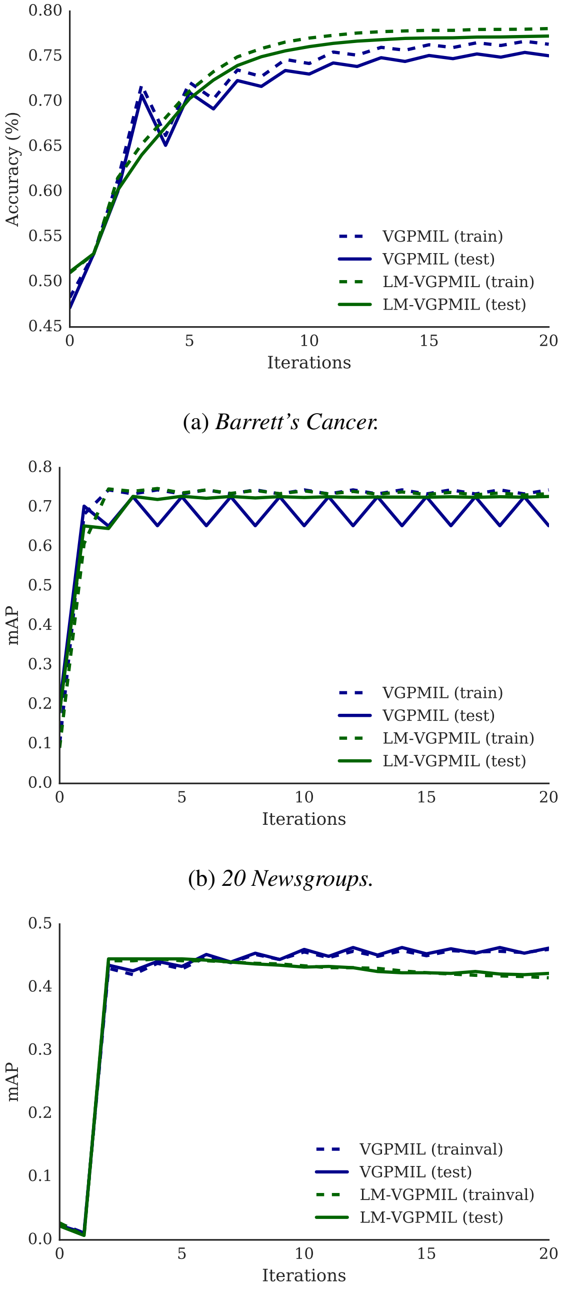


(a) *t-SNE on Barrett's cancer data set.*



(b) *t-SNE on PASCAL VOC 2007.*

Figure 1: t-SNE [8] visualization. For the Barrett's cancer data set all of the instances coloured according to the correct instance labels are plotted. For PASCAL VOC 2007 we plotted the (for the model unavailable) ground truth instances.

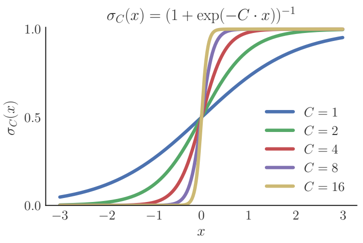


(a) *Barrett's Cancer.*

(b) *20 Newsgroups.*

(c) *PASCAL VOC 2007.*

Figure 2: **Learning Curves.** Both models have very steep learning curves and tend to come very close to their final performance after just a couple of iterations. While the regularization introduced in the LM-VGPMIL leads to smoother learning curves in all three data sets it's overall benefit varies between them. While it clearly outperforms VGPMIL in the Barrett's Cancer data set, for the 20 Newsgroups data base the curves are a lot closer. Both models come close to their final performance after just three iterations, but while the large margin model can deal with the heavy imbalance in the bags and their cluttered features, the base model keeps oscillating (as stated in the main text, this behaviour can be alleviated by further preprocessing the features *e.g.* with a kernel PCA). Figure (c) shows a case where the additional regularization actually hurts the predictive performance. While LM-VGPMIL initially reaches a higher performance than the VGPMIL model, its performance decreases with further iterations, whereas VGPMIL improves further.



(a) Effect of C

V	$\sigma(V)$
0	0.5
0.5	0.62
1	0.73
1.5	0.82
2	0.88
2.5	0.92
3	0.95

(b) Effect of V

Figure 3: Influence of C and V on the logistic sigmoid. With larger values of C, the sigmoid approximates the step function more closely and penalizes margin violations more heavily. Similarly, V shifts the sigmoid along the x-axis.

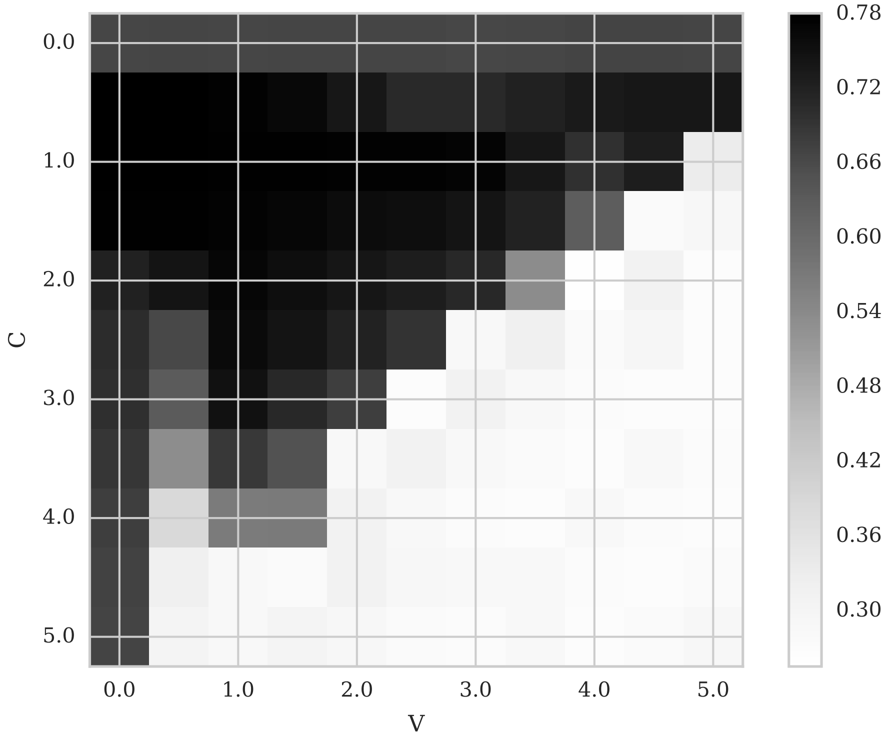


Figure 4: Influence of C and V in the LM-VGPMIL model trained on the Barrett's Cancer dataset.

References

[1] H. Bilen and A. Vedaldi. Weakly supervised deep detection networks. In *CVPR*, 2016. 3

[2] R. G. Cinbis, J. Verbeek, and C. Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016. 3

[3] M. Kandemir and F. A. Hamprecht. Instance label prediction by Dirichlet process multiple instance learning. In *UAI*, 2014. 4

[4] V. Kantorov, M. Oquab, M. Cho, and I. Laptev. Contextlocnet: Context-aware deep network models for weakly supervised localization. In *ECCV*, 2016. 3

[5] D. Kotzias, M. Denil, N. De Freitas, and P. Smyth. From group to individual labels using deep features. In *SIGKDD*. ACM, 2015. 4

[6] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang. Weakly supervised object localization with progressive domain adaptation. In *CVPR*, 2016. 3

[7] G. Liu, J. Wu, and Z.-H. Zhou. Key instance detection in multi-instance learning. *ACML*, 2012. 4

[8] L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov), 2008. 5

[9] M. Shi and V. Ferrari. Weakly supervised object localization using size estimates. In *ECCV*, 2016. 3

[10] P. Spirtes. Directed cyclic graphical representations of feedback models. In *UAI*, 1995. 2

[11] E. W. Teh, M. Rochan, and Y. Wang. Attention networks for weakly supervised object localization. In *BMVC*, 2016. 3