Supplementary Materials: Active Learning for Structured Probabilistic Models with Histogram Approximation

In this supplementary material, we present the proofs of Lemma 1 and Lemma 2.

1. Proof of Lemma 1

Lemma 1. *Let* $Q(\mathbf{y};\mathbf{Y},\mathbf{q}) = \sum_{m=1}^{M} q_m [\![\mathbf{y} = \mathbf{y}^m]\!]$ *be a SOWD-approximation parameterized by* $\mathbf{Y}$ *and* $\mathbf{q}$*. Let* $KL(Q||\mathrm{P}) = \sum_{\mathbf{y}\in\mathcal{Y}} Q(\mathbf{y}) \log \frac{Q(\mathbf{y})}{\mathrm{P}(\mathbf{y}|\mathbf{x})}$ *denote the KL-divergence between the two distributions. The parameters* $\hat{\mathbf{Y}}, \hat{\mathbf{q}}$ *that minimize* $KL(Q||\mathrm{P})$ *are:*

$$\hat{\mathbf{y}}^m = \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} \ \mathrm{P}(\mathbf{y}|\mathbf{x}) \tag{1a}$$

$$s.t. \ \ \mathbf{y} \neq \hat{\mathbf{y}}^{m'} \quad \forall m' < m \tag{1b}$$

$$\hat{q}_m = \frac{e^{S(\hat{\mathbf{y}}^m)}}{\sum_{m'=1}^{M} e^{S(\hat{\mathbf{y}}^{m'})}} \tag{2}$$

Proof.

$$KL(Q||\mathrm{P}) = \sum_{\mathbf{y}\in\mathcal{Y}} Q(\mathbf{y}) \log \frac{Q(\mathbf{y})}{\mathrm{P}(\mathbf{y}|\mathbf{x})} \tag{3a}$$

$$= \sum_{\mathbf{y}\in\mathbf{Y}} Q(\mathbf{y}) \log \frac{Q(\mathbf{y})}{\mathrm{P}(\mathbf{y}|\mathbf{x})} + \sum_{\mathbf{y}\in\mathcal{Y}\setminus\mathbf{Y}} Q(\mathbf{y}|\mathbf{x}) \log \frac{Q(\mathbf{y})}{\mathrm{P}(\mathbf{y}|\mathbf{x})} \tag{3b}$$

$$= \sum_{\mathbf{y}\in\mathbf{Y}} Q(\mathbf{y}) \log \frac{Q(\mathbf{y})}{\mathrm{P}(\mathbf{y}|\mathbf{x})} \tag{3c}$$

Thus, we have

$$\min_{\mathbf{Y},\mathbf{q}} KL(Q||\mathrm{P}) \Rightarrow \begin{cases} \min\limits_{q_m,\mathbf{y}^m} & \sum\limits_{m=1}^{M} q_m \log \frac{q_m}{\mathrm{P}(\mathbf{y}^m|\mathbf{x})} \\ s.t. & \sum\limits_{m=1}^{M} q_m = 1 \end{cases} \tag{4}$$

We can write the Lagrangian for Eqn. (4) as

$$L(\mathbf{Y},\mathbf{q},\lambda) = \sum_{m=1}^{M} q_m \log \frac{q_m}{\mathrm{P}(\mathbf{y}^m|\mathbf{x})} + \lambda \cdot \left(\sum_{m=1}^{M} q_m - 1 \right) \tag{5}$$

Method of Lagrangian multipliers involves setting the derivative of L w.r.t q_m to 0,

$$\frac{\partial L}{\partial q_m} = \log \frac{q_m}{\mathrm{P}(\mathbf{y}^m|\mathbf{x})} + 1 + \lambda = 0 \tag{6}$$

Thus,

$$q_m = e^{-1-\lambda} \mathrm{P}(\mathbf{y}^m|\mathbf{x}) \tag{7}$$

Using the fact that $\sum\limits_{m=1}^{M} q_m = 1$, we can show that $\lambda = \log \left(\sum\limits_{m=1}^{M} \mathrm{P}(\mathbf{y}^m|\mathbf{x}) \right) - 1$. Thus,

$$q_m = \frac{\mathrm{P}(\mathbf{y}^m|\mathbf{x})}{\sum_{m'=1}^{M} \mathrm{P}(\mathbf{y}^{m'}|\mathbf{x})} \tag{8}$$

Plugging this definition of q_m in objective function of Eqn. (4), we get

$$\sum_{m=1}^{M} q_m \log \frac{q_m}{\mathrm{P}(\mathbf{y}^m|\mathbf{x})} \tag{9}$$

$$= \sum_{m=1}^{M} \frac{\mathrm{P}(\mathbf{y}^m|\mathbf{x})}{\sum_{m'=1}^{M} \mathrm{P}(\mathbf{y}^{m'}|\mathbf{x})} \log \frac{\frac{\mathrm{P}(\mathbf{y}^m|\mathbf{x})}{\sum_{m'=1}^{M} \mathrm{P}(\mathbf{y}^{m'}|\mathbf{x})}}{\mathrm{P}(\mathbf{y}^m|\mathbf{x})} \tag{10}$$

$$= \sum_{m=1}^{M} \frac{\mathrm{P}(\mathbf{y}^m|\mathbf{x})}{\sum_{m'=1}^{M} \mathrm{P}(\mathbf{y}^{m'}|\mathbf{x})} \log \frac{1}{\sum_{m'=1}^{M} \mathrm{P}(\mathbf{y}^{m'}|\mathbf{x})} \tag{11}$$

$$= \left(\sum_{m=1}^{M} \mathrm{P}(\mathbf{y}^m|\mathbf{x}) \right) \cdot \left(\frac{1}{\sum_{m'=1}^{M} \mathrm{P}(\mathbf{y}^{m'}|\mathbf{x})} \log \frac{1}{\sum_{m'=1}^{M} \mathrm{P}(\mathbf{y}^{m'}|\mathbf{x})} \right) \tag{12}$$

$$= -\log \sum_{m'=1}^{M} \mathrm{P}(\mathbf{y}^{m'}|\mathbf{x}) \tag{13}$$

Thus,

$$\min_{\mathbf{Y},\mathbf{q}} KL(Q||\mathrm{P}) \Rightarrow \max_{\mathbf{Y}} \sum_{\mathbf{y}\in\mathbf{Y}} \mathrm{P}(\mathbf{y}|\mathbf{x}) \tag{14}$$

Clearly, $\sum\limits_{\mathbf{y}\in\mathbf{Y}} \mathrm{P}(\mathbf{y}|\mathbf{x})$ is maximized by picking the top M probability locations in P,

$$\hat{\mathbf{y}}^m = \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} \ \mathrm{P}(\mathbf{y}|\mathbf{x}) \tag{15}$$

$$s.t. \ \ \mathbf{y} \neq \hat{\mathbf{y}}^{m'} \quad \forall m' < m \tag{16}$$

978-1-4673-6964-0/15/$31.00 ©2015 IEEE

$$\hat{q}_m = \frac{\mathrm{P}(\hat{\mathbf{y}}^m|\mathbf{x})}{\sum_{m'=1}^{M}\mathrm{P}(\hat{\mathbf{y}}^{m'}|\mathbf{x})} = \frac{e^{S(\hat{\mathbf{y}}^m)}}{\sum_{m'=1}^{M} e^{S(\hat{\mathbf{y}}^{m'})}} \tag{17}$$

This completes the proof. We can see that the optimal Q is a normalized distribution over the top M most probable locations in P. □

2. Proof of Lemma 2

Lemma 2. *Let* $Q(\mathbf{y}; \{\mathcal{Y}^m\}, \mathbf{q}) = \sum_{m=1}^{M} \frac{q_m}{|\mathcal{Y}^m|}[\![\mathbf{y} \in \mathcal{Y}^m]\!]$ *be a histogram-approximation parameterized by bins* $\{\mathcal{Y}^m\}$ *and weights* $\mathbf{q}$. *Let* $KL(\mathrm{P}||Q) = \sum_{\mathbf{y}\in\mathcal{Y}} \mathrm{P}(\mathbf{y}|\mathbf{x}) \log \frac{\mathrm{P}(\mathbf{y}|\mathbf{x})}{Q(\mathbf{y})}$ *denote the KL-divergence between the two distributions. For any fixed set of non-overlapping (potentially unequally sized) bins* $\{\mathcal{Y}^m\}$, *such that* $\mathcal{Y} = \cup_m \mathcal{Y}^m$, *the weights* $\hat{\mathbf{q}}$ *that minimize* $KL(\mathrm{P}||Q)$ *are:*

$$\hat{q}_m = \sum_{\mathbf{y}\in\mathcal{Y}^m} \mathrm{P}(\mathbf{y}|\mathbf{x}) = \frac{1}{\mathcal{Z}} \sum_{\mathbf{y}\in\mathcal{Y}^m} e^{S(\mathbf{y})} \tag{18}$$

Proof.

$$KL(\mathrm{P}||Q) = \sum_{m=1}^{M} \sum_{\mathbf{y}\in\mathcal{Y}^m} \mathrm{P}(\mathbf{y}|\mathbf{x}) \log \left(\frac{\mathrm{P}(\mathbf{y}|\mathbf{x})}{\frac{q_m}{|\mathcal{Y}^m|}} \right) \tag{19}$$

$$= \sum_{m=1}^{M} \sum_{\mathbf{y}\in\mathcal{Y}^m} \mathrm{P}(\mathbf{y}|\mathbf{x}) \log \left(\mathrm{P}(\mathbf{y}|\mathbf{x}) \cdot |\mathcal{Y}^m|\right) - \sum_{m=1}^{M} \sum_{\mathbf{y}\in\mathcal{Y}^m} \mathrm{P}(\mathbf{y}|\mathbf{x}) \log q_m \tag{20}$$

$$= h - \sum_{m=1}^{M} \sum_{\mathbf{y}\in\mathcal{Y}^m} \mathrm{P}(\mathbf{y}|\mathbf{x}) \log q_m \tag{21}$$

$$= h - \sum_{m=1}^{M} p_m \log q_m \tag{22}$$

where, h is a constant and $p_m = \sum_{\mathbf{y}\in\mathcal{Y}^m} \mathrm{P}(\mathbf{y}|\mathbf{x})$, i.e., the mass of the distribution in the bin m. Thus,

$$\min_{\mathbf{q}} KL(\mathrm{P}||Q) \Rightarrow \begin{cases} \min\limits_{q_m} & -\sum\limits_{m=1}^{M} p_m \log q_m \\ s.t., & \sum\limits_{m=1}^{M} q_m = 1 \end{cases} \tag{23}$$

We can write the Lagrangian for Eqn. (23) as

$$L(\mathbf{q}, \lambda) = -\sum_{m=1}^{M} p_m \log q_m + \lambda \left(\sum_{m=1}^{M} q_m - 1 \right) \tag{24}$$

Method of Lagrangian multipliers involves setting the derivative of L w.r.t q_m to 0,

$$\frac{\partial L}{\partial q_m} = -\frac{p_m}{q_m} + \lambda = 0 \Rightarrow q_m = \frac{p_m}{\lambda} \tag{25}$$

Using the fact that $\sum_{m=1}^{M} q_m = \sum_{m=1}^{M} \frac{p_m}{\lambda} = \frac{1}{\lambda} = 1$, we can show that $\lambda = 1$. Thus, $\hat{q}_m = p_m$, i.e.,

$$\hat{q}_m = \sum_{\mathbf{y}\in\mathcal{Y}^m} \mathrm{P}(\mathbf{y}|\mathbf{x}) = \frac{1}{\mathcal{Z}} \sum_{\mathbf{y}\in\mathcal{Y}^m} e^{S(\mathbf{y})} \tag{26}$$

This completes the proof. We can see that the optimal Q is a normalized histogram over M bins.

□

3. Qualitative Results

Fig. 1,2 show example images with the most uncertainty/certainty according to our approach `Active-PDivMAP` and `Gibbs`, under a model trained with 5 images from 2 random trails in the experiment section of the main paper.

We can see that `Gibbs` has difficulty transitioning out of one mode to reach another mode. As a result, almost all sampled segmentations look visually very similar, and the estimated distribution/histogram is nearly uniform. From these two examples, we can see that `Gibbs` will typically pick images where the MAP is already pretty accurate – the model will seem uncertain because `Gibbs` is picking samples that are all very similar to MAP.

In contrast, our approach `Active-PDivMAP` can pick images (first row) for which the set of plausible segmentations (or histogram bin centers) are truly diverse, but have similar energies. Such images are much more helpful in updating the beliefs of the model in an active learning setting.

Note that in both examples, our approach estimates the entropy of the most uncertain image to be ≈ 2.29 (compared to the maximum possible entropy of a 10-D probability mass function $\log 10 = 2.30$).

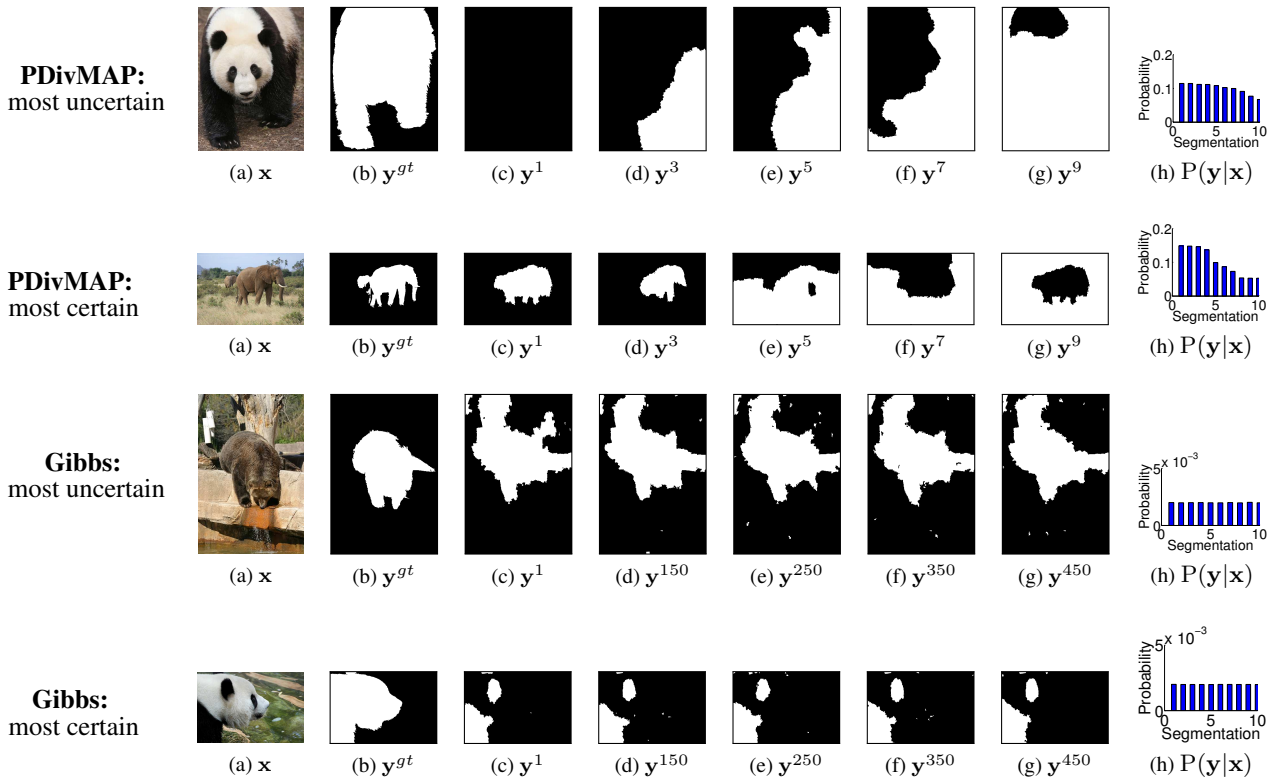

Figure 1: Example 1: First/second row shows the image with the most uncertainty/certainty, as estimated by our approach `Active-PDivMAP`. Third/fourth row shows the image with the most uncertainty/certainty, as estimated by `Gibbs`. We can see that `Gibbs` has difficulty transitioning out of one mode to reach another mode. Thus, almost all sampled segmentations of the most uncertain image look visually very similar. In contrast, our approach `Active-PDivMAP` can pick images (first row) for which the set of plausible segmentations (or histogram bin centers) are truly diverse, but have similar energies. This identifies images where the model is truly uncertain, and such images are helpful in updating the beliefs of the model.

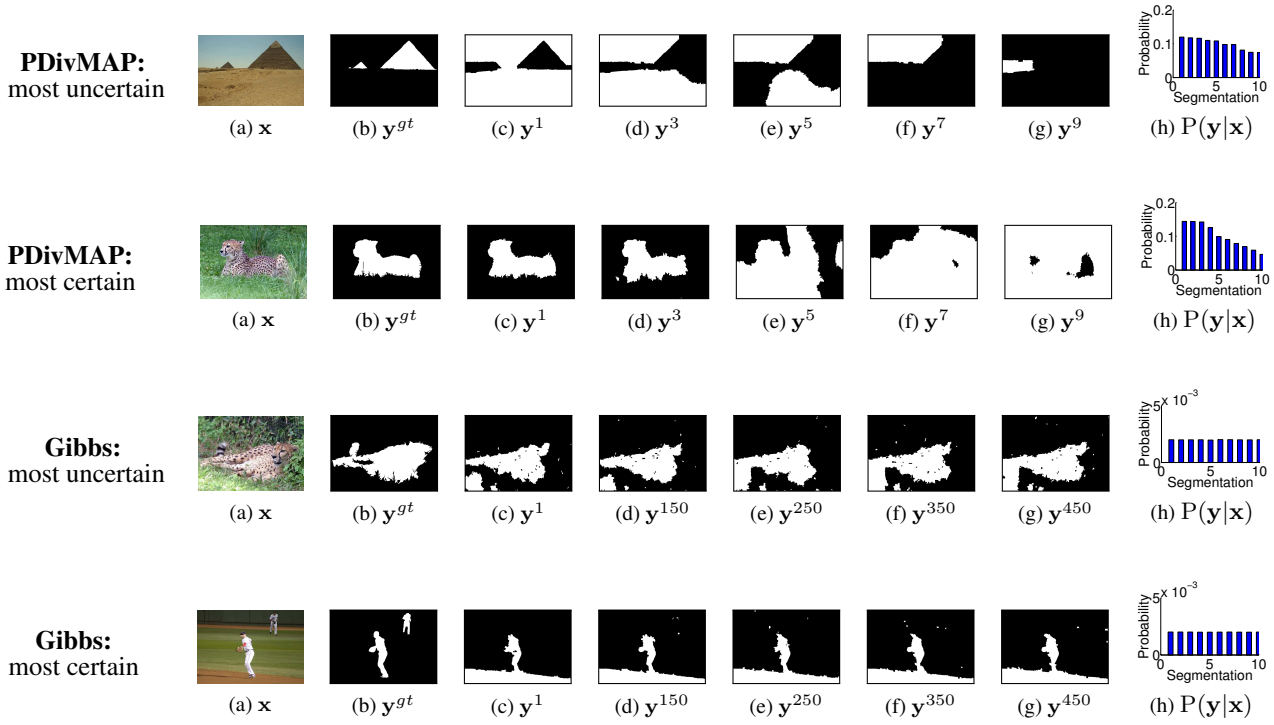

Figure 2: Example 2: First/second row shows the image with the most uncertainty/certainty, as estimated by our approach `Active-PDivMAP`. Third/fourth row shows the image with the most uncertainty/certainty, as estimated by `Gibbs`. We can see that `Gibbs` has difficulty transitioning out of one mode to reach another mode. Thus, almost all sampled segmentations of the most uncertain image look visually very similar. In contrast, our approach `Active-PDivMAP` can pick images (first row) for which the set of plausible segmentations (or histogram bin centers) are truly diverse, but have similar energies. This identifies images where the model is truly uncertain, and such images are helpful in updating the beliefs of the model.