BAYESIAN FEW-SHOT CLASSIFICATION WITH ONE-VS-EACH PÓLYA-GAMMA AUGMENTED GAUSSIAN PROCESSES: ADDITIONAL EXPERIMENTS

Anonymous authors
Paper under double-blind review

1 LIKELIHOOD VISUALIZATION

In order to visualize the various likelihoods under consideration, we first consider a trivial classification task with a single observed example. We assume that there are three classes ($C = 3$) and the single example belongs to the first class ($y = 1$). We place the following prior on $\mathbf{f} = (f_1, f_2, f_3)^\top$:

$$p(\mathbf{f}) = \mathcal{N}\left(\mathbf{f} \,\middle|\, \boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}\right). \tag{1}$$

In other words, the prior for f_1 and f_2 is a standard normal and f_3 is clamped at zero (for ease of visualization). We consider here the softmax, Gaussian, logistic softmax (LSM), and one-vs-each (OVE) likelihoods as defined in Table 1.

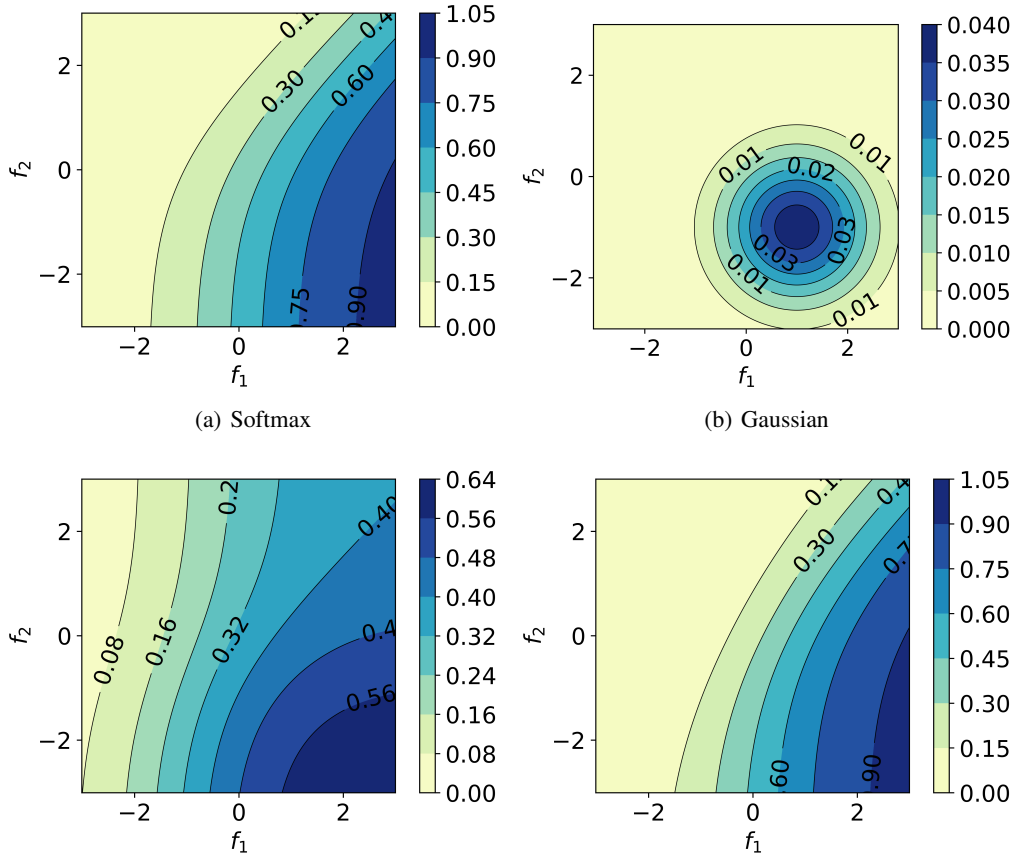
Table 1: Definition of likelihoods for Section 1.

Likelihood	$L(\mathbf{f}\|y=1)$
Softmax	$\frac{e^{f_1}}{e^{f_1}+e^{f_2}+e^{f3}}$
Gaussian	$\mathcal{N}(1 \,\|\, \mu = f_1, \sigma^2 = 1)\mathcal{N}(-1 \,\|\, \mu = f_2, \sigma^2 = 1)\mathcal{N}(-1 \,\|\, \mu = f_3, \sigma^2 = 1)$
Logistic Softmax (LSM)	$\frac{\sigma(f_1)}{\sigma(f_1)+\sigma(f_2)+\sigma(f_3)}$
One-vs-Each (OVE)	$\sigma(f_1 - f_2)\sigma(f_1 - f_3)$

The posterior for a given likelihood can be computed as:

$$p(\mathbf{f}|y=1) = \frac{L(\mathbf{f}|y=1)p(\mathbf{f})}{\int L(\mathbf{f}'|y=1)p(\mathbf{f}')\,d\mathbf{f}'} \tag{2}$$

The likelihoods as defined in Table 1 and corresponding posteriors are plotted in Figure 1 and Figure 2, respectively.



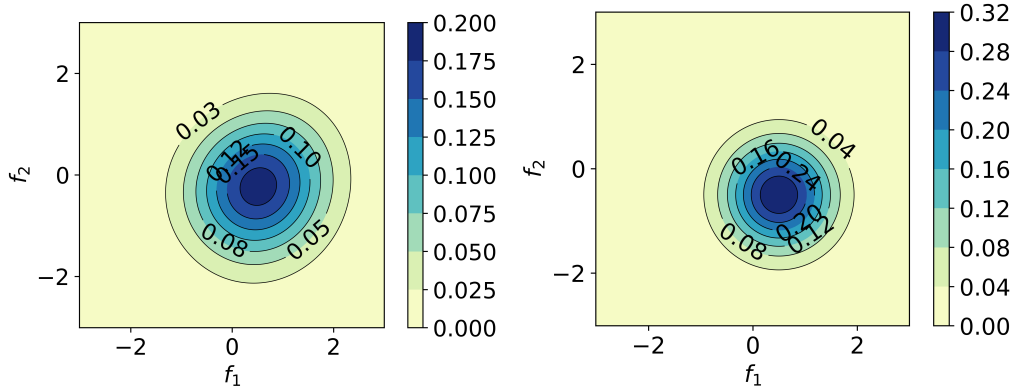
(a) Softmax

(b) Gaussian

(c) Logistic Softmax

(d) One-vs-Each

Figure 1: Plot of $L(\mathbf{f}|y = 1)$, where f_3 is clamped to 0. Gaussian likelihood penalizes configurations far away from $(f_1, f_2) = (1, -1)$. Logistic softmax is much flatter compared to softmax and has visibly different contours. One-vs-Each is visually similar to the softmax but penalizes (f_1, f_2) near the origin slightly more.



(a) Softmax

(b) Gaussian

(c) Logistic Softmax

(d) One-vs-Each

Figure 2: Plot of posterior $p(\mathbf{f}|y = 1)$, where f_3 is clamped to 0. The mode of each posterior distribution is similar, but each differs slightly in shape. Gaussian is more peaked about its mode, while logistic softmax is more spread out. One-vs-Each is similar to softmax, but is slightly more elliptical.

2 2D IRIS EXPERIMENTS

We also conducted experiments on a 2D version of the Iris dataset, which contains 150 examples across 3 classes. The first two features of the dataset were retained (sepal length and width). We used a zero-mean GP prior and an RBF kernel $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2}d(\mathbf{x}, \mathbf{x}'\right)^2$, where $d(\cdot, \cdot)$ is Euclidean distance. We considered class-balanced training set sizes with 1, 2, 3, 4, 5, 10, 15, 20, 25, and 30 examples per class. For each training set size, we performed GP inference on 200 randomly generated train/test splits and compared the predictions across Gaussian, logistic softmax, and one-vs-each likelihoods.

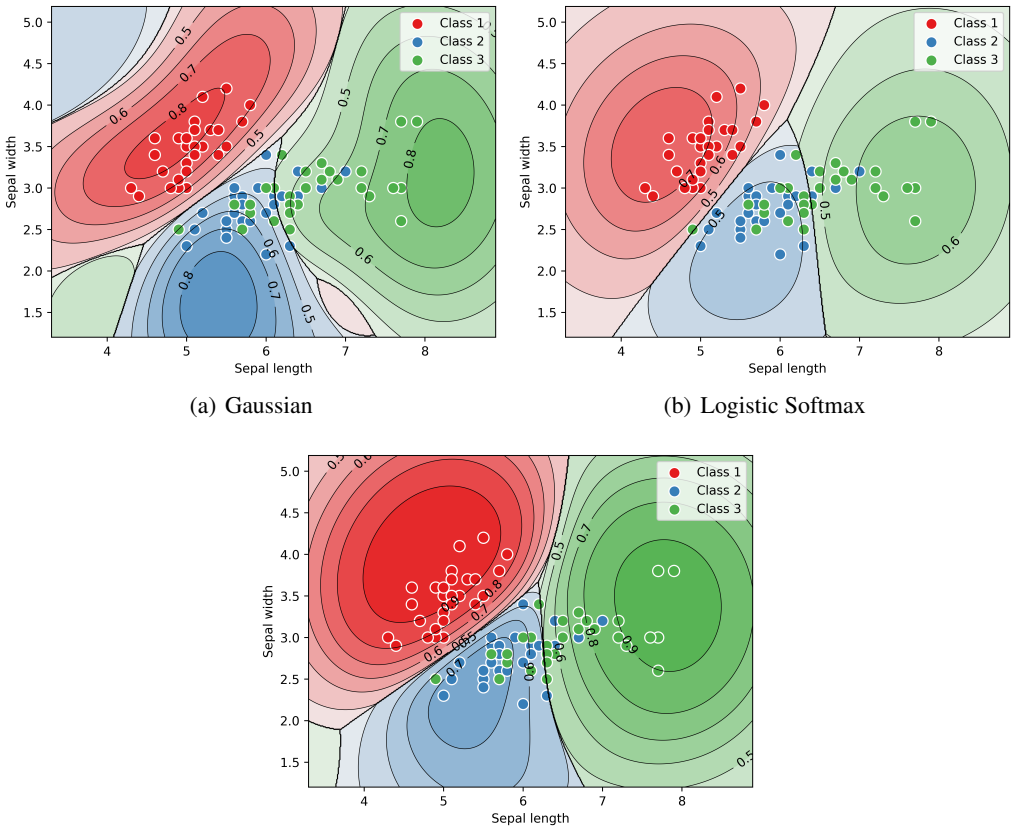
Predictions at a test point $\mathbf{x}_*$ were made by applying the normalized likelihood to the posterior predictive mean $\bar{\mathbf{f}}_*$. The predictive probabilities for each likelihood is shown in Figure 3 for a randomly generated train/test split with 30 examples per class. Test predictive accuracy, Brier score, expected calibration error, and evidence lower bound (ELBO) results across various training set sizes are shown in Figure 4.

The ELBO is computed by treating each likelihood's posterior $q(\mathbf{f}|\mathbf{X}, \mathbf{Y})$ as an approximation to the softmax posterior $p(\mathbf{f}|\mathbf{X}, \mathbf{Y})$.

$$\begin{aligned}\text{ELBO}(q) &= \mathbb{E}_q[\log p(\mathbf{f}|\mathbf{X})] + \mathbb{E}_q[\log p(\mathbf{Y}|\mathbf{f})] - \mathbb{E}_q[\log q(\mathbf{f}|\mathbf{X}, \mathbf{Y})] \\ &= \log p(\mathbf{x}) - \text{KL}(q(\mathbf{f}|\mathbf{X}, \mathbf{Y})||p(\mathbf{f}|\mathbf{X}, \mathbf{Y})).\end{aligned}$$

Even though direct computation of the softmax posterior $p(\mathbf{f}|\mathbf{X}, \mathbf{y})$ is intractable, computing the ELBO is tractable. A larger ELBO indicates a lower the KL divergence to the softmax posterior.

One-vs-Each performs well for accuracy, Brier score, and ELBO across the training set sizes. Gaussian performs best on expected calibration error through 15 examples per class, beyond which one-vs-each is better.

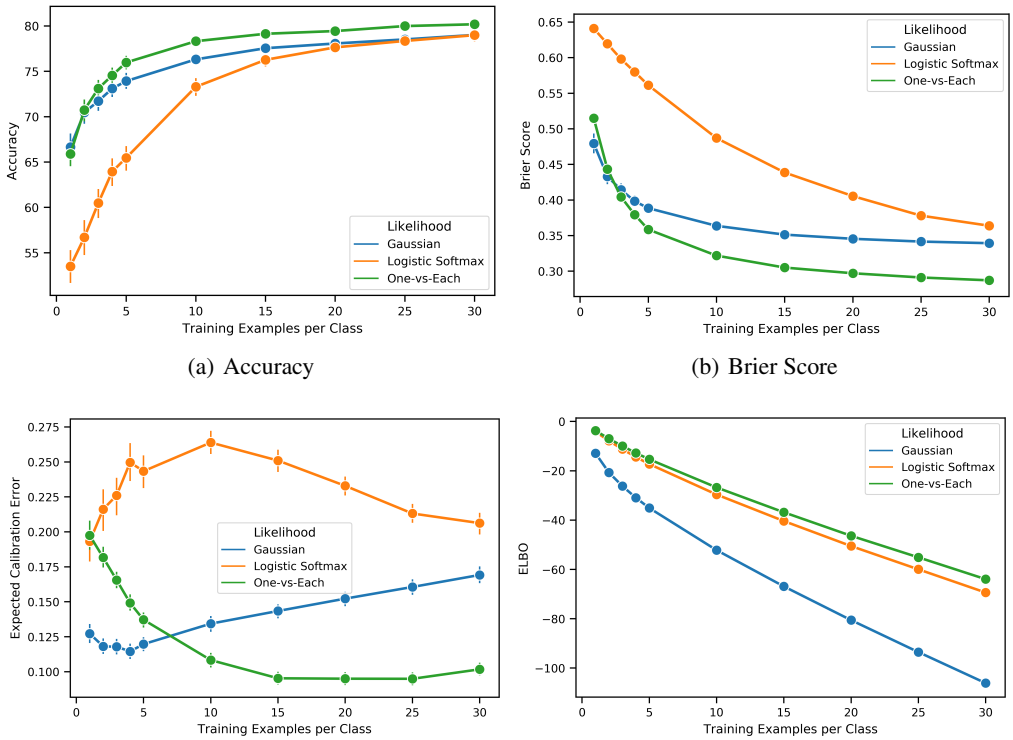


(a) Gaussian

(b) Logistic Softmax

(c) One-vs-Each

Figure 3: Training points (colored points) and maximum predictive probability for various likelihoods on the Iris dataset. The Gaussian likelihood produces more warped decision boundaries than the others. Logistic softmax tends to produce lower confidence predictions, while one-vs-each produces larger regions of greater confidence than the others.



(a) Accuracy

(b) Brier Score

(c) Expected Calibration Error (computed with 10 bins)

(d) ELBO

Figure 4: Comparison across likelihoods in terms of test predictive accuracy, Brier score, and expected calibration error Results are averaged over 200 randomly generated splits for each training set size (1, 2, 3, 4, 5, 10, 15, 20, 25, and 30 examples per class). Error bars indicate 95% confidence intervals.

3 A CLOSER LOOK AT THE ONE-VS-EACH LIKELIHOOD

In this section, we examine the form of the one-vs-each posterior compared to the softmax posterior. The posterior $p(\mathbf{f}|\mathbf{X}, \mathbf{y})$ can be written in terms of the unnormalized posterior $g(\mathbf{f}|\mathbf{X}, \mathbf{y})$:

$$p(\mathbf{f}|\mathbf{X}, \mathbf{y}) = \frac{g(\mathbf{f}|\mathbf{X}, \mathbf{y})}{Z(\mathbf{X}, \mathbf{y})}, \text{ where} \tag{3}$$

$$Z(\mathbf{X}, \mathbf{y}) = \int g(\mathbf{f}|\mathbf{X}, \mathbf{y})\, d\mathbf{f}, \text{ and} \tag{4}$$

$$g(\mathbf{f}|\mathbf{X}, \mathbf{y}) = p(\mathbf{f}|\mathbf{X}) \prod_{i=1}^{N} \mathcal{L}(\mathbf{f}_i|y_i). \tag{5}$$

We now turn our attention to the forms of the one-vs-each and softmax likelihoods, denoted by $L^{\text{OVE}}(\mathbf{f}_i|y_i = c)$ and $L^{\text{SM}}(\mathbf{f}_i|y_i = c)$, respectively. The one-vs-each likelihood is a lower bound on the softmax likelihood, as shown by (Titsias, 2016):

$$\begin{aligned} L^{\text{SM}}(\mathbf{f}_i|y_i = c) &= \frac{1}{1 + \sum_{c' \neq c} e^{-(f_i^c - f_i^{c'})}} \\ &\geq \prod_{c' \neq c} \frac{1}{1 + e^{-(f_i^c - f_i^{c'})}} = \prod_{c' \neq c} \sigma(f_i^c - f_i^{c'}) = L^{\text{OVE}}(\mathbf{f}_i|y_i = c). \end{aligned} \tag{6}$$

This allows us to conclude that the one-vs-each unnormalized posterior and normalizing constant are both lower bounds on the corresponding softmax quantities:

$$g^{\text{OVE}}(\mathbf{f}|\mathbf{X}, \mathbf{y}) = p(\mathbf{f}|\mathbf{X}) \prod_{i=1}^{N} \mathcal{L}^{\text{OVE}}(\mathbf{f}_i|y_i) \leq p(\mathbf{f}|\mathbf{X}) \prod_{i=1}^{N} \mathcal{L}^{\text{SM}}(\mathbf{f}_i|y_i) = g^{\text{SM}}(\mathbf{f}|\mathbf{X}, \mathbf{y}) \tag{7}$$

$$Z^{\text{OVE}}(\mathbf{X}, \mathbf{y}) = \int g^{\text{OVE}}(\mathbf{f}|\mathbf{X}, \mathbf{y}) \leq \int g^{\text{SM}}(\mathbf{f}|\mathbf{X}, \mathbf{y}) = Z^{\text{SM}}(\mathbf{X}, \mathbf{y}) \tag{8}$$

We can interpret the one-vs-each posterior as "upweighting" some regions of $\mathbf{f}$, for which $p^{\text{OVE}}(\mathbf{f}|\mathbf{X}, \mathbf{y}) > p^{\text{SM}}(\mathbf{f}|\mathbf{X}, \mathbf{y})$, and "downweighting" others, for which $p^{\text{OVE}}(\mathbf{f}|\mathbf{X}, \mathbf{y}) < p^{\text{SM}}(\mathbf{f}|\mathbf{X}, \mathbf{y})$. In particular,

$$p^{\text{OVE}}(\mathbf{f}|\mathbf{X}, \mathbf{y}) < p^{\text{SM}}(\mathbf{f}|\mathbf{X}, \mathbf{y}) \Rightarrow \frac{g^{\text{OVE}}(\mathbf{f}|\mathbf{X}, \mathbf{y})}{g^{\text{SM}}(\mathbf{f}|\mathbf{X}, \mathbf{y})} < \frac{Z^{\text{OVE}}(\mathbf{X}, \mathbf{y})}{Z^{\text{SM}}(\mathbf{X}, \mathbf{y})}. \tag{9}$$

In other words, if the ratio of unnormalized densities evaluated at $\mathbf{f}$ is less than the ratio of normalizing constants (which is simply the unnormalized density integrated over $\mathbf{f}$), then the posterior density for OVE will be lower. We can gain intuition about where those values are likely to occur by examining the bound equation 6. When the bound is tight, the OVE posterior is likely to have higher density than the softmax posterior. When the bound is relatively loose, the reverse is likely to happen.

Consider as an illustrative example the case where $C = 4$ and $y_i = 1$. The softmax likelihood is then:

$$L^{\text{SM}}(\mathbf{f}_i|y_i = 1) = \frac{e^{f_i^1}}{e^{f_i^1} + e^{f_i^2} + e^{f_i^3} + e^{f_i^4}}, \tag{10}$$

where f_i^c denotes the function value for class c and data point i. The corresponding one-vs-each likelihood is:

$$\begin{aligned} L^{\text{OVE}}(\mathbf{f}_i|y_i = 1) &= \left(\frac{e^{f_i^1}}{e^{f_i^1} + e^{f_i^2}}\right)\left(\frac{e^{f_i^1}}{e^{f_i^1} + e^{f_i^3}}\right)\left(\frac{e^{f_i^1}}{e^{f_i^1} + e^{f_i^4}}\right) \\ &= \frac{e^{f_i^1}}{e^{f_i^1} + e^{f_i^2} + e^{f_i^3} + e^{f_i^4} + e^{f_i^2+f_i^3-f_i^1} + e^{f_i^2+f_i^4-f_i^1} + e^{f_i^3+f_i^4-f_i^1} + e^{f_i^2+f_i^3+f_i^4-2f_i^1}}. \end{aligned} \tag{11}$$

Whether the bound is tight or loose will depend on the extra terms present in the denominator of equation 11. Comparing the form of equation 10 and equation 11, we can conclude that the bound will be tight if f_i^1 is much larger than $f_i^{c'}$ for each $c' \in \{2, 3, 4\}$, because the $-f_i^1$ and $-2f_i^1$ will dominate the corresponding terms. If $f_i^{c'}$ is similar to or larger than f_i^1 for some $c' \in \{2, 3, 4\}$, then the bound will be loose, implying that the posterior density will be lower for OVE than the softmax. However, if $f_i^{c'}$ is *much* larger than f_i^1 for some $c' \in \{2, 3, 4\}$, the overall likelihood will be low and thus have limited impact on the posterior.

We can observe this phenomenon in Figure 2. When compared to the softmax posterior, the OVE posterior looks "squashed" towards the lower right corner, which is where the ground truth function value (f_1) is much greater than the non-ground truth function value (f_2). The upper left corner, where f_2 is greater than f_1, has less posterior density compared to the softmax, as we would expect given the above discussion.

REFERENCES

Michalis K Titsias. One-vs-each approximation to softmax for scalable estimation of probabilities. *Advances in Neural Information Processing Systems*, 29:4161–4169, 2016.