Let i and o denote that a sample was drawn from the in- or out-distributions respectively. Then the model is constructed as follows:

$$
\begin{aligned}
p(y|x) &= p(y,i|x) + p(y,o|x) && (1)\\
&= p(y|i,x)p(i|x) + p(y|o,x)p(o|x) && (2)\\
&= p(y|i,x)\frac{p(x|i)p(i)}{p(x|i)p(i)+p(x|o)p(o)} + \frac{1}{M}\frac{p(x|o)p(o)}{p(x|i)p(i)+p(x|o)p(o)} && (3)\\
&= \frac{p(y|i,x)p(x|i) + \frac{\lambda}{M}p(x|o)}{p(x|i)+\lambda p(x|o)} && (4)
\end{aligned}
$$

where $\lambda \equiv \frac{p(o)}{p(i)}$. We also use that $p(y|o,x) = p(y|o) = \frac{1}{M}$, where M is the number of classes. If on our input domain $\mathcal{D} = [0,1]^d$ we want to be agnostic about the out-distribution, we may assume $p(x|o) = 1$. We also note that the formulation still works if $p(y|i,x)$ is not a probability distribution but instead is simply non-negative and with finite Lebesque-measure over the domain, because the corresponding normalization factor can simply be reabsorbed into λ. These considerations motivate the study of the following theorem:

Theorem 1. *Let*

$$p(y|x) = \frac{p(y|i,x)p(x|i) + \frac{\lambda}{M}}{p(x|i)+\lambda}$$

on the domain $\mathcal{D} = [0,1]^d$, where $0 \leq p(y|i,x) \leq 1$, $\lambda > 0$ and $M > 1$. Assume that on the domain $p(x|i)$ is given as

$$p(x|i) = \sum_{k=0}^{K} \alpha_k \exp\left(-\frac{d(x,\mu_k)}{2\sigma_k^2}\right)$$

with $\alpha_k > 0$, $\sigma_k^2 > 0$ and $\mu_k \in \mathcal{D} \quad \forall k = 1,\dots K$ and $d(\cdot,\cdot) : \mathcal{D}\times\mathcal{D} \to \mathbb{R}^+$ a metric. Furthermore, define $L_k \equiv \min_n d(x_n,\mu_k)$, where all x_n are from some finite subset $X \subset \mathcal{D}$. Then

$$\forall \epsilon > \frac{1}{M}\ \exists L \in \mathbb{R}_+\ \forall z \in \mathcal{D} : \min_n d(x_n,z) \geq L \implies p(y|x) \leq \epsilon$$

Proof. The proof works by constructing an explicit bound for any z satisfying $\min_n d(x_n,z) \geq L$ and then showing that it can be made to be arbitrarily close to $\frac{1}{M}$ by increasing L. First notice that $p(y|i,x) \leq 1$ implies

$$
\begin{aligned}
p(y|x) &\leq \frac{p(x|i)+\frac{\lambda}{M}}{p(x|i)+\lambda} && (5)\\
&= \frac{1}{M}\frac{1+M\frac{p(x|i)}{\lambda}}{1+\frac{p(x|i)}{\lambda}} && (6)\\
&\leq \frac{1}{M}\frac{1+M\frac{b}{\lambda}}{1+\frac{b}{\lambda}} \quad \forall b \geq p(x|i) && (7)
\end{aligned}
$$

The last step holds because the function $\frac{1+M\xi}{1+\xi}$ is monotonically increasing in ξ for $\xi > 0$ and $M > 1$, because

$$\frac{d}{d\xi}\frac{1+M\xi}{1+\xi} = \frac{M-1}{(1+\xi)^2} > 0 \quad \forall M > 1, \xi > 0 \tag{8}$$

and since $\frac{p(x|i)}{\lambda} > 0$ this yields a bound, as long as we can indeed find some $b \geq p(z|i)$. Starting from the definition of $p(z|i)$ we see that

$$p(z|i) = \sum_{k=1}^{K} \alpha_k \exp\left(-\frac{d(z,\mu_k)^2}{2\sigma_k^2}\right) \tag{9}$$

$$\leq \sum_{k=1}^{K} \alpha_k \exp\left(-\frac{D_k^2}{2\sigma_k^2}\right) \tag{10}$$

where we introduce a set of D_k that must satisfy $D_k \leq d\left(z,\mu_k\right)$. But such a set of numbers can easily be found via the reverse triangle inequality if we assume that $L > L_k \quad \forall k$

$$d(z,\mu_k) \geq |d(z,x_n) - d(x_n,\mu_k)| \tag{11}$$

$$\geq L - L_k \equiv D_k \tag{12}$$

Thus, we have found a valid bound for

$$p(y|x) \leq \frac{1}{M}\frac{1+M\frac{b}{\lambda}}{1+\frac{b}{\lambda}} \tag{13}$$

$$b = \sum_{k=1}^{K} \alpha_k \exp\left(-\frac{(L-L_k)^2}{2\sigma_k^2}\right) \tag{14}$$

By choosing L sufficiently large we can get b arbitrarily close to zero and, in turn, the bound on $p(y|x)$ will get arbitrarily close to $\frac{1}{M}$.

□

Note that the above theorem can easily be extended to arbitrary mixture models of the form

$$p(x|i) = \sum_{k} \alpha_k g\left(d(\mu_k, x)\right) \tag{15}$$

where $g : \mathbb{R}^+ \to \mathbb{R}^+$ has to be both bounded and monotonically decreasing.

We can get a sense for how tight the bound is by exploring it on MNIST, as shown in 1. The bound clearly follows the actual shape of the data samplesbut is not tight. However, by looking at the second row of plots we can appreciate that the bound is actually useful because it does take care of the overconfidence we have on far-away samples.

While Theorem 1 is true, it would be useful to have an explicit construction that allows us to infer a required distance L from the training set in order to reach low confidence rather than the other way around. We achieve this with the following theorem. The assumptions are almost identical to the previous one, except now we simplify things by putting the centroids on datapoints.

Theorem 2. *Let*

$$p(y|x) = \frac{p(y|i,x)p(x|i) + \frac{\lambda}{M}}{p(x|i) + \lambda}$$

on the domain $\mathcal{D} = [0,1]^d$*, where* $0 \leq p(y|i,x) \leq 1$*,* $\lambda > 0$ *and* $M > 1$*. Assume that on the domain* $p(x|i)$ *is given as*

$$p(x|i) = \sum_{k=0}^{K} \alpha_k \exp\left(-\frac{d(x,\mu_k)}{2\sigma_k^2}\right)$$

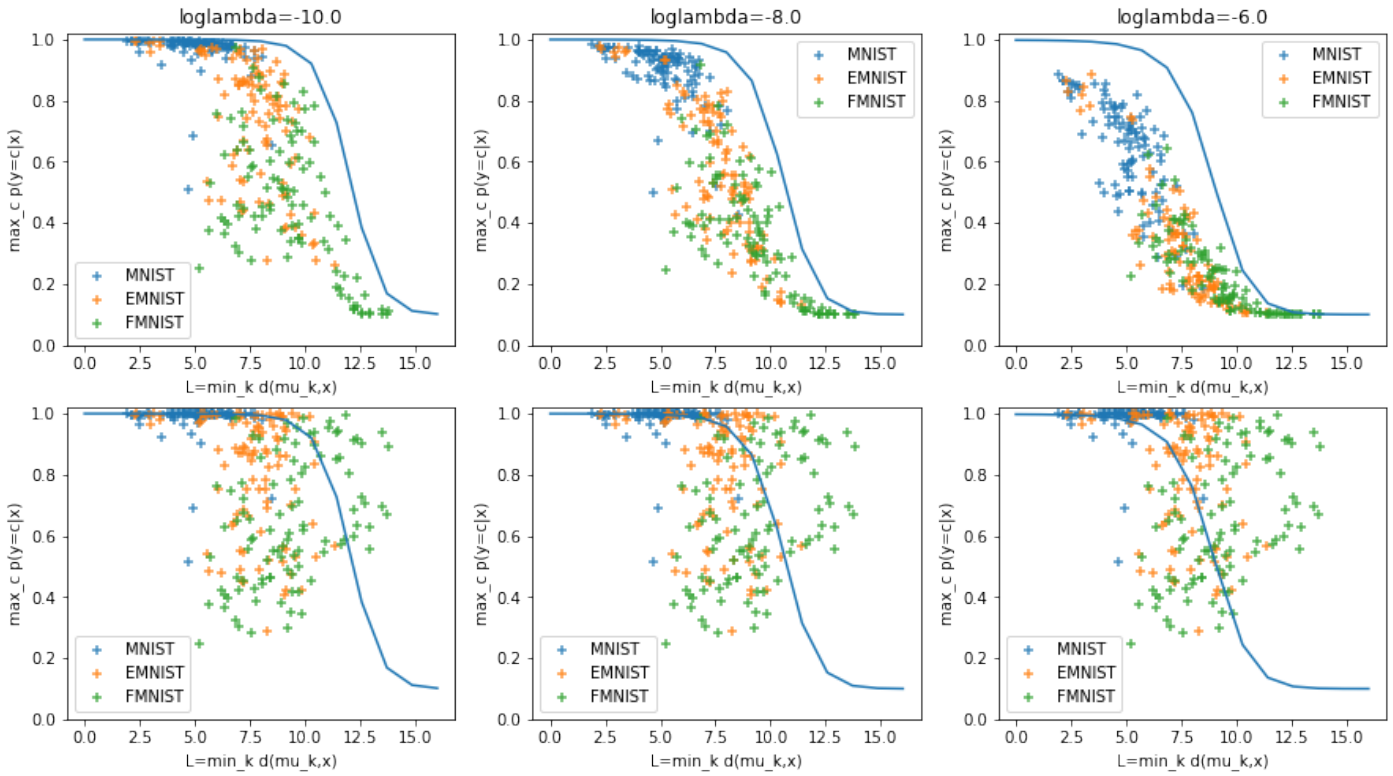

Figure 1: Bounds on maximum confidence as a function of the distance to the closest data point on a restricted subset of the training data for different values of λ. The centroids are located on points of this subset.

with $\alpha_k > 0$, $\sigma_k^2 > 0$ *and* $\mu_k \in \mathcal{D} \quad \forall k = 1, \dots K$ *and* $d(\cdot, \cdot) : \mathcal{D} \to [0, 1]$ *a metric. Furthermore let* $X \subset \mathcal{D}$ *and assume that* $\{\mu_1, \dots, \mu_K\} \subset X$. *Then for any* $\epsilon > 0$ *we can*

$$\forall \epsilon > 0 \forall z \in \mathcal{D} : \left(\min_{x \in X} d(z, x) = L \wedge L^2 \geq 2 \max_k \left(\sigma_k^2\right) \log \left[\frac{(M-1) \sum_k \alpha_k}{\lambda \epsilon} \right] \right) \implies p(y|x) \leq \frac{1}{M}(1 + \epsilon)$$

Proof. We can immediately see from Theorem 1 that a valid L must exist that would satisfy the inequality. The only new part will be the explicit sufficient condition that L must satisfy. We recycle from the previous proof the result

$$p(y|x) \leq \frac{1}{M} \frac{1 + M\frac{b}{\lambda}}{1 + \frac{b}{\lambda}} \quad \forall b \geq p(x|i) \tag{16}$$

$$b = \sum_{k=1}^{K} \alpha_k \exp\left(-\frac{L^2}{2\sigma_k^2}\right) \tag{17}$$

In order to write this in an adequate format for the current theorem we need an additional inequality

$$\frac{1 + M\frac{b}{\lambda}}{1 + \frac{b}{\lambda}} \leq 1 + (M-1)\frac{b}{\lambda} \tag{18}$$

This can easily be seen by noticing that $\frac{1+M\xi}{1+\xi}$ is concave down, i.e.

$$\frac{d^2}{d\xi^2} \frac{1 + M\xi}{1 + \xi} = -2\frac{M-1}{(1+\xi)^3} < 0 \quad \forall M > 1, \xi > 0 \tag{19}$$

so we can be sure that for $\xi > 0$ a first order Taylor-expansion will overestimate the function. We thus find

$$p(y|x) \leq \frac{1}{M}\left(1 + (M-1)\frac{b}{\lambda}\right) \tag{20}$$

$$= \frac{1}{M}\left(1 + \frac{M-1}{\lambda}\sum_{k=1}^{K}\alpha_k \exp\left(-\frac{L^2}{2\sigma_k^2}\right)\right) \tag{21}$$

$$\leq \frac{1}{M}\left(1 + \frac{M-1}{\lambda}\left(\sum_{k=1}^{K}\alpha_k\right)\exp\left(-\frac{L^2}{2\max_k \sigma_k^2}\right)\right) \tag{22}$$

$$\overset{!}{\leq} \frac{1}{M}(1+\epsilon) \tag{23}$$

By rearranging the last inequality we find the condition under which it applies, namely

$$L^2 \geq 2\max_k\left(\sigma_k^2\right)\log\left[\frac{(M-1)\sum_k \alpha_k}{\lambda\ \epsilon}\right]. \tag{24}$$

□