Technical Report

Deep Evidential Epistemic and Aleatoric Uncertainty Estimation for Semantic Segmentation in Off-Road Navigation

SECTION I
BAYESIAN UNCERTAINTY DECOMPOSITION

It is important to distinguish between *aleatoric* and *epistemic* uncertainty [21, 30, 43]. *Aleatoric* uncertainty arises from inherent and irreducible ambiguity in the true class label due to noise in sensor observations such as motion blur or low-resolution of small objects. Aleatoric uncertainty can be computed from the softmax output of the classifier and trained by maximizing data log-likelihood. On the other hand, *epistemic* uncertainty arises from uncertainty in model parameters due limited training data (when multiple models can explain/fit the same training dataset). Epistemic uncertainty is directly related to *out-of-distribution* detection; epistemic uncertainty is high for examples that lie outside the model's training distribution. Conveniently, in the Bayesian estimation framework, the total uncertainty i.e. entropy of the label posterior distribution can be decomposed into two terms that correspond to aleatoric and epistemic uncertainty, as derived below. A graphical model for the framework is depicted in

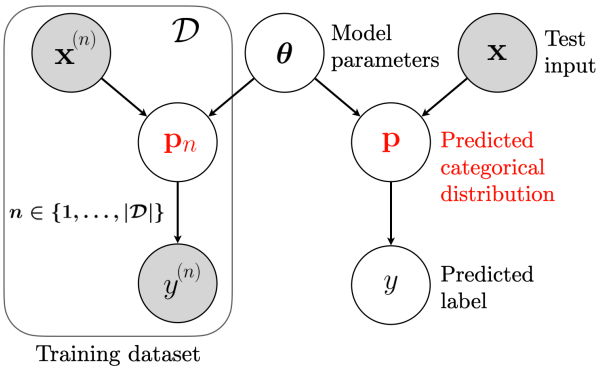


Fig. 4: Probabilistic graphical model of Bayesian uncertainty estimation in classification/segmentation. A classifier model parametrized by $\boldsymbol{\theta}$ outputs a categorical probability distribution $\mathbf{p}$ given an input $\mathbf{x}$. The predicted label y is then sampled from $\mathbf{p}$. Given a dataset $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^{|\mathcal{D}|}$, the task is to infer the distributional posterior $p(\mathbf{p} \,|\, \mathbf{x}, \mathcal{D})$ for test input $\mathbf{x}$.

Fig. 4. A single classifier (a.k.a. "model") is parametrized by $\boldsymbol{\theta} \in \Theta$, where Θ is the space of model parameters. Given a D-dimensional input vector $\mathbf{x} \in \mathbb{R}^D$, each model $\boldsymbol{\theta}$ predicts a categorical distribution $\mathbf{p} = (p_1, p_2, \ldots, p_C) \in \Delta^C$ over C discrete classes where $\mathbf{p}$ lies in the C-dimensional simplex i.e. $\forall c,\ p_c \in [0, 1]$ and $\sum_{c=1}^{C} p_c = 1$. The predicted semantic label $y \in \{1, \ldots, C\}$ is then sampled from $\mathbf{p}$.

Given a training dataset $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}_{i=1}^{|\mathcal{D}|}$ of input-label pairs $(\mathbf{x}^{(n)}, y^{(n)})$, Bayesian neural networks learn to predict the *label posterior* distribution (also called the "posterior predictive" distribution) for a given test input $\mathbf{x}$ by marginalizing out the model parameters $\boldsymbol{\theta}$ and categorical distributions $\mathbf{p}$:

$$\underbrace{p(y \,|\, \mathbf{x}, \mathcal{D})}_{\text{label posterior}} = \int \int_{\boldsymbol{\theta}, \mathbf{p}} p(y \mid \mathbf{p})\ p(\mathbf{p} \mid \mathbf{x}, \boldsymbol{\theta})\ p(\boldsymbol{\theta} \mid \mathcal{D})\ \mathrm{d}\mathbf{p}\ \mathrm{d}\boldsymbol{\theta} \tag{8}$$

Throughout the paper, we will treat the term $p(\mathbf{p} \mid \mathbf{x}, \boldsymbol{\theta})$ as point mass: a classifier $\boldsymbol{\theta}$ deterministically predicts a single categorical distribution $\mathbf{p}$ for a given input $\mathbf{x}$. However, Eqn. 8 is more general and makes mathematical notation convenient [37]. Classical formulations of Bayesian uncertainty marginalize $\mathbf{p}$ in Eqn. 8 to obtain $p(y \mid \mathbf{x}, \mathcal{D}) = \int_{\boldsymbol{\theta}} p(y \mid \mathbf{x}, \boldsymbol{\theta})\ p(\boldsymbol{\theta} \mid \mathcal{D})\ \mathrm{d}\boldsymbol{\theta}$ and focus on the *model posterior* $p(\boldsymbol{\theta} \mid \mathcal{D})$. However, to study epistemic and aleatoric uncertainty, it is instead more useful to marginalize out $\boldsymbol{\theta}$ from Eqn. 8 to obtain:

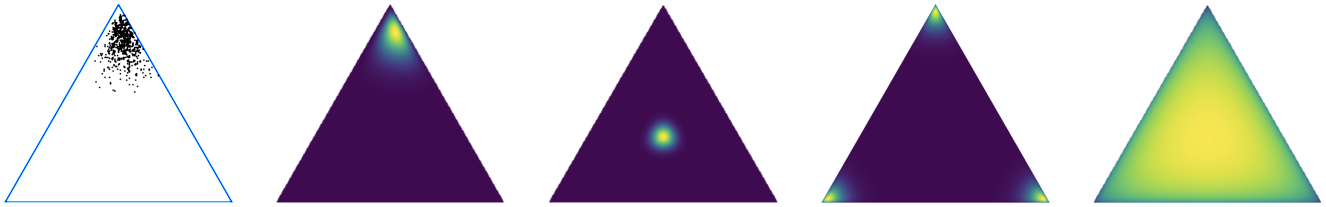
$$\underbrace{p(y \,|\, \mathbf{x}, \mathcal{D})}_{\overline{\mathbf{p}}\text{: label posterior}} = \int_{\mathbf{p}} p(y \,|\, \mathbf{p})\ \underbrace{p(\mathbf{p} \,|\, \mathbf{x}, \mathcal{D})}_{\pi\text{: distributional posterior}}\ \mathrm{d}\mathbf{p} \tag{9}$$

where $\pi \coloneqq p(\mathbf{p} \,|\, \mathbf{x}, \mathcal{D}) = \int_{\boldsymbol{\theta}} p(\mathbf{p} \mid \mathbf{x}, \boldsymbol{\theta})\ p(\boldsymbol{\theta} \mid \mathcal{D})\ \mathrm{d}\boldsymbol{\theta}$ is known as the *distributional posterior*. The distributional posterior shall be our main quantity of interest. It is a posterior distribution over categorical distributions $\mathbf{p}$. For example, a trained ensemble of k models predicts k categorical distributions $\pi = (\mathbf{p}_1, \ldots, \mathbf{p}_k)$ for an input $\mathbf{x}$. Note that the final label posterior distribution $\overline{\mathbf{p}} \coloneqq p(y \,|\, \mathbf{x}, \mathcal{D})$ is simply the mean of the distributional posterior π.

$$\overline{\mathbf{p}} = \mathbb{E}_{\mathbf{p} \sim \pi} \mathbf{p} \tag{10}$$

Therefore the final label distribution of the ensemble would be $\overline{\mathbf{p}} = \frac{1}{k} \sum_{j=1}^{k} \mathbf{p}_j$. Let us denote $\overline{\mathbf{p}} = (\overline{p}_1, \ldots, \overline{p}_C)$, where $\overline{p}_c = \mathbb{E}_{\mathbf{p} \sim \pi}\ p_c$. The uncertainty of a categorical distribution $\mathbf{p}$ is measured by its *entropy*: $\mathbb{H}(\mathbf{p}) = -\sum_{c=1}^{C} p_c \log p_c$ [12]. Then, the total uncertainty (entropy) of the label posterior distribution (final output distribution over labels) $\mathbb{H}(\overline{\mathbf{p}})$ can be decomposed as [15, 37]:

$$\begin{aligned}
\underbrace{\mathbb{H}(\overline{\mathbf{p}})}_{\text{Total uncertainty}} &= -\sum_{c=1}^{C} \overline{p}_c \log \overline{p}_c \\
&= -\sum_{c=1}^{C} \left(\mathbb{E}_{\mathbf{p} \sim \pi}\ p_c\right) \log \overline{p}_c \\
&= -\mathbb{E}_{\mathbf{p} \sim \pi} \sum_{c=1}^{C} p_c \log \overline{p}_c \\
&= -\mathbb{E}_{\mathbf{p} \sim \pi} \sum_{c=1}^{C} p_c \left[\log \frac{\overline{p}_c}{p_c} + \log p_c \right]
\end{aligned}$$



(a) Finite mixture of categorical distributions predicted by members of an ensemble.

(b) Corresponding Dirichlet posterior; **low**-epistemic and **low**-aleatoric uncertainty.

(c) Dirichlet posterior with **low**-epistemic and **high**-aleatoric uncertainty.

(d) Dirichlet posterior with **high**-epistemic and **low**-aleatoric uncertainty.

(e) Dirichlet posterior with **high**-epistemic and **high**-aleatoric uncertainty.

Fig. 5: Visualization of different types of distributional posteriors $p(\mathbf{p} \mid \mathbf{x}, \mathcal{D})$ for three semantic classes. Each triangle denotes the three-dimensional simplex Δ^3; each point on the simplex is a categorical distribution $\mathbf{p} \in \Delta^3$ with $p_1, p_2, p_3 \in [0, 1]$ and $p_1 + p_2 + p_3 = 1$. **(a)** Distributional posterior from an ensemble, which is a finite mixture of categorical distributions; each categorical is predicted by a member of the ensemble. **(b)** A Dirichlet posterior over Δ^3 that approximates the ensemble distribution. **(b-d)** Various types of Dirichlet posteriors. Distributions that are concentrated at a point (vs. spread out) have low (vs. high) epistemic uncertainty since all categorical samples are in agreement. Distributions that are close to the corners of the simplex (vs. close to the center) have low (vs. high) aleatoric uncertainty since all categorical samples are more (vs. less) certain.

$$
\begin{aligned}
&= -\mathbb{E}_{\mathbf{p}\sim\pi} \sum_{c=1}^{C} p_c \log \frac{\overline{p}_c}{p_c} - \mathbb{E}_{\mathbf{p}\sim\pi} \sum_{c=1}^{C} p_c \log p_c \\
\underbrace{\mathbb{H}(\overline{\mathbf{p}})}_{\text{Total uncertainty}} &= \underbrace{\mathbb{E}_{\mathbf{p}\sim\pi} \mathbb{D}_{\mathrm{KL}}(\mathbf{p} \,||\, \overline{\mathbf{p}})}_{\text{Epistemic uncertainty}} + \underbrace{\mathbb{E}_{\mathbf{p}\sim\pi} \mathbb{H}(\mathbf{p})}_{\text{Aleatoric uncertainty}} \qquad (11) \\
&\text{where } \overline{\mathbf{p}} \coloneqq \mathbb{E}_{\mathbf{p}\sim\pi} \mathbf{p}
\end{aligned}
$$

The epistemic uncertainty is the mean distance (KL-divergence) between the predicted categoricals and the mean categorical. In other words, it measures the *disagreement* between the predicted categoricals across multiple models $p(\boldsymbol{\theta} \mid \mathcal{D})$; the disagreement is expected to be high for OOD test inputs $\mathbf{x}$ that are not well represented in the training data. The aleatoric uncertainty is the average entropy of the predicted categoricals; aleatoric uncertainty is high when most models $p(\boldsymbol{\theta} \mid \mathcal{D})$ predict a uniform distribution over labels due to noisy inputs $\mathbf{x}$. Note that each of the three uncertainties in Eqn. 11 lie in $[0, \log C]$.

Fig. 5 visualizes different types of distributional posteriors that are parametrized as Dirichlet distributions, and their implications for epistemic and aleatoric uncertainty.

Section II
Theoretical Analysis of the Corrected Bayesian Loss Function

Posterior networks [6, 7] are a recently proposed evidential deep learning method that predict $p(\mathbf{p} \mid \mathbf{x}, \mathcal{D})$ motivated by exact posterior inference for the Dirichlet-categorical conjugate pair. In Sec. II-A, we describe the exact Bayesian inference procedure for the Dirichlet-categorical conjugate pair in an idealistic scenario. Then, in Sec. II-B, we detail the use of NatPNs to approximate the exact posterior inference for Dirichlet distributions.

A. Exact posterior inference for Dirichlet distributions

Consider a idealistic scenario where the test input $\mathbf{x}$ (or examples similar to $\mathbf{x}$) are observed in the training dataset $\mathcal{D}$, $N^{\mathbf{x}}$ times: $\mathcal{D} = \{(\mathbf{x}, y^{(n)})\}_{n=1}^{N^{\mathbf{x}}}$ and the label samples $y^{(n)}$ are i.i.d. from $p(y \mid \mathbf{x})$. A large $N^{\mathbf{x}}$ should correspond to low epistemic uncertainty and a high $N^{\mathbf{x}}$ should correspond to high epistemic uncertainty. Before observing any labeled data, Charpentier et al. [6, 7] assume $p(\mathbf{p} \mid \mathbf{x}) = \mathrm{Dir}(\mathbf{p} \mid \mathbf{1})$, where $\mathbf{1} = (1, ..., 1)$ is a C-dimensional vector. $\mathrm{Dir}(\mathbf{p} \mid \mathbf{1})$ is a flat prior whose probability density is uniform over its support $\mathbf{p} \in \Delta^C$. After observing the labels $\{y^{(n)}\}_{n=1}^{N^{\mathbf{x}}}$, the posterior distribution is also a Dirichlet (due to conjugacy) and is given by $p(\mathbf{p} \mid \mathbf{x}, \mathcal{D}) = \mathrm{Dir}(\mathbf{p} \mid \mathbf{1} + N^{\mathbf{x}} \boldsymbol{\beta}^{\mathbf{x}})$ [58], where $\boldsymbol{\beta}^{\mathbf{x}}$ is a C-dimensional vector whose c-th component is the fraction of observed labels with class c i.e.

$$
N_c^{\mathbf{x}} \coloneqq |\{y^{(n)} \mid y^{(n)} = c\}_{n=1}^{N^{\mathbf{x}}}| \qquad \beta_c^{\mathbf{x}} \coloneqq \frac{N_c^{\mathbf{x}}}{N^{\mathbf{x}}} \qquad (12)
$$

Note that $N^{\mathbf{x}}$ corresponds to the "count" of $\mathbf{x}$ in the training set i.e. the number of examples in the training set "resemble" $\mathbf{x}$. This is also referred to as the "pseudo-count" of $\mathbf{x}$. The pseudo-count is approximated by $N^{\mathbf{x}} \approx N^{\boldsymbol{\phi}}(\mathbf{x}) = N_H p_{\boldsymbol{\theta}}(\mathbf{x})$ where $p_{\boldsymbol{\theta}}(\mathbf{x})$ is a probability density estimate of $\mathbf{x}$ from an invertible normalizing flow model with weights $\boldsymbol{\phi}$, and N_H is a constant scaling factor. Similarly $\boldsymbol{\beta}^{\mathbf{x}}$ is the observed empirical distribution of labels of $\mathbf{x}$ in $\mathcal{D}$.

B. Approximating posterior Dirichlet inference via NatPNs

Charpentier et al. [6, 7] approximate the exact posterior distribution $\mathrm{Dir}(\mathbf{p} \mid \mathbf{1} + N^{\mathbf{x}} \boldsymbol{\beta}^{\mathbf{x}})$ described in App. II-B by combining the outputs of a classification network (in our case, a segmentation network) $\boldsymbol{\beta}^{\boldsymbol{\phi}}(\mathbf{x}) \in \Delta^C$, and a density estimator (normalizing flow network) $p_{\boldsymbol{\theta}}(\mathbf{x}) \in \mathbb{R}_{\geq 0}$. The classifier $\boldsymbol{\beta}^{\boldsymbol{\phi}}(\mathbf{x})$ outputs a categorical distribution over labels given $\mathbf{x}$, whereas the scaled output of the density estimator $N^{\boldsymbol{\phi}}(\mathbf{x}) = N_H \; p_{\boldsymbol{\theta}}(\mathbf{x})$ is used to approximate the pseudo-count of $\mathbf{x}$. The overall neural-network based approximation is given by:

$$
\begin{aligned}
p(\mathbf{p} \mid \mathbf{x}, \mathcal{D}) &= \mathrm{Dir}(\mathbf{p} \mid \mathbf{1} + N^{\mathbf{x}} \boldsymbol{\beta}^{\mathbf{x}}) \\
&\approx \mathrm{Dir}(\mathbf{p} \mid \mathbf{1} + \underbrace{N^{\boldsymbol{\phi}}(\mathbf{x})}_{=N_H \; p_{\boldsymbol{\theta}}(\mathbf{x})} \cdot \boldsymbol{\beta}^{\boldsymbol{\phi}}(\mathbf{x}))
\end{aligned}
$$

The weights $\boldsymbol{\phi}$ of the normalizing flow model $p_{\boldsymbol{\theta}}$ and the classifier $\boldsymbol{\beta}^{\boldsymbol{\phi}}$ are shared since they both share a learned latent representation [7].

Going forward, we drop the dependence on $\mathbf{x}$ for notational simplicity i.e. write $\pi_{\boldsymbol{\phi}}(\mathbf{x})$ as $\pi_{\boldsymbol{\phi}}$, $N^{\boldsymbol{\phi}}(\mathbf{x})$ as $N^{\boldsymbol{\phi}}$ and $\boldsymbol{\beta}^{\boldsymbol{\phi}}(\mathbf{x})$ as $\boldsymbol{\beta}^{\boldsymbol{\phi}}$. Furthermore, we define

$$\boldsymbol{\alpha}^{\mathbf{x}} := \mathbf{1} + N^{\mathbf{x}}\boldsymbol{\beta}^{\mathbf{x}} \quad \text{(true posterior Dirichlet parameters)}$$
$$\boldsymbol{\alpha}^{\boldsymbol{\phi}} := \mathbf{1} + N^{\boldsymbol{\phi}}\boldsymbol{\beta}^{\boldsymbol{\phi}} \quad \text{(predicted posterior Dirichlet parameters)}$$

where $\boldsymbol{\alpha}^{\mathbf{x}}, \boldsymbol{\alpha}^{\boldsymbol{\phi}} \in \{\mathbb{R}_{>0}\}^C$ are C-dimensional positive vectors.

C. Deriving the coefficient for the Bayesian loss function using variational inference

The evidential neural network with parameters $\boldsymbol{\phi}$ learns to predict the posterior distribution $\mathrm{Dir}(\mathbf{p} \mid \mathbf{1} + N^{\mathbf{x}}\boldsymbol{\beta}^{\mathbf{x}}) \approx \pi_{\boldsymbol{\phi}}(\mathbf{x}) = \mathrm{Dir}(\mathbf{p} \mid \mathbf{1} + N^{\boldsymbol{\phi}}(\mathbf{x})\boldsymbol{\beta}^{\boldsymbol{\phi}}(\mathbf{x}))$ by minimizing a "Bayesian loss function" [6, 7]. The loss combines an expected cross-entropy term that encourages the classifier to predict the conditional label distribution $p(y \mid \mathbf{x})$, with an entropy regularization term that prevents the predicted Dirichlet distribution from being excessively concentrated/peaky.

$$\mathcal{L}(\boldsymbol{\phi}) = \sum_n \Big[\underbrace{-\mathbb{E}_{\mathbf{p}\sim\pi_{\boldsymbol{\phi}}} \log(y^{(n)} \mid \mathbf{p})}_{\text{Expected cross-entropy loss term}} - \lambda \underbrace{\mathbb{H}(\pi_{\boldsymbol{\phi}})}_{\text{Entropy regularization term}} \Big]$$
$$\text{where } \pi_{\boldsymbol{\phi}} = \mathrm{Dir}(\mathbf{p} \mid \mathbf{1} + N^{\boldsymbol{\phi}}\boldsymbol{\beta}^{\boldsymbol{\phi}}) \tag{13}$$

Since $\pi_{\boldsymbol{\phi}}(\mathbf{x})$ is a Dirichlet distribution, both terms of the Bayesian loss function (Eqn. 7) can be computed in closed-form and are differentiable [6, 7].

The combination coefficient λ balances between training the classifier to minimize cross-entropy loss, with an entropy regularization term that affects the density estimator (invertible normalizing flow) and controls the concentration of the predicted Dirichlet. Charpentier et al. [6, 7] hand-tune λ and set it to a constant value.

Can the value of λ be *derived* in a principled way? We want the predicted posterior distribution $\pi_{\boldsymbol{\phi}}(\mathbf{p}) = \mathrm{Dir}(\mathbf{p} \mid \mathbf{1} + N^{\boldsymbol{\phi}}\boldsymbol{\beta}^{\boldsymbol{\phi}})$ to approximate the true posterior $p(\mathbf{p} \mid \mathcal{D}) = \mathrm{Dir}(\mathbf{p} \mid \mathbf{1} + N^{\mathbf{x}}\boldsymbol{\beta}^{\mathbf{x}})$, where $\mathcal{D} = \{y^{(n)}\}_{n=1}^{N^{\mathbf{x}}}$ and $\beta_c^{\mathbf{x}} = |\{y^{(n)} \mid y^{(n)} = c\}_{n=1}^{N^{\mathbf{x}}}|/N^{\mathbf{x}}$. Therefore, we pose the problem as variational inference where we want to minimize the KL-divergence between a parametric predicted posterior distribution $\pi_{\boldsymbol{\phi}}(\mathbf{p})$ and the true posterior distribution $p(\mathbf{p} \mid \mathcal{D})$:

$$\begin{aligned}
&\mathbb{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\phi}}(\mathbf{p}) \,||\, p(\mathbf{p} \mid \mathcal{D})) \\
&= -\mathbb{E}_{\mathbf{p}\sim\pi_{\boldsymbol{\phi}}} \log \frac{p\big(\mathbf{p} \mid \{y^{(n)}\}_{n=1}^{N^{\mathbf{x}}}\big)}{\pi_{\boldsymbol{\phi}}(\mathbf{p})} \\
&= -\mathbb{E}_{\mathbf{p}\sim\pi_{\boldsymbol{\phi}}} \log \frac{p\big(\{y^{(n)}\}_{n=1}^{N^{\mathbf{x}}} \mid \mathbf{p}\big)\ p(\mathbf{p})}{\pi_{\boldsymbol{\phi}}(\mathbf{p})} \\
&= -\mathbb{E}_{\mathbf{p}\sim\pi_{\boldsymbol{\phi}}} \log p\big(\{y^{(n)}\}_{n=1}^{N^{\mathbf{x}}} \mid \mathbf{p}\big) + \mathbb{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\phi}}(\mathbf{p}) \,||\, p(\mathbf{p})) \\
&= \Big[-\mathbb{E}_{\mathbf{p}\sim\pi_{\boldsymbol{\phi}}} \log \prod_{n=1}^{N^{\mathbf{x}}} p(y^{(n)} \mid \mathbf{p}) \Big] + \mathbb{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\phi}}(\mathbf{p}) \,||\, \mathrm{Dir}(\mathbf{p} \mid \mathbf{1})) \\
&= \Big[\sum_{n=1}^{N^{\mathbf{x}}} -\mathbb{E}_{\mathbf{p}\sim\pi_{\boldsymbol{\phi}}} \log p(y^{(n)} \mid \mathbf{p}) \Big] - \mathbb{H}(\pi_{\boldsymbol{\phi}}(\mathbf{p})) + \text{constant} \\
&\equiv \sum_{n=1}^{N^{\mathbf{x}}} \Big[-\mathbb{E}_{\mathbf{p}\sim\pi_{\boldsymbol{\phi}}} \log p(y^{(n)} \mid \mathbf{p}) - \frac{1}{N^{\mathbf{x}}}\, \mathbb{H}(\pi_{\boldsymbol{\phi}}(\mathbf{p})) \Big]
\end{aligned} \tag{14}$$

Here, the prior $p(\mathbf{p} \mid \mathbf{1})$ is the uniform Dirichlet distribution $\mathrm{Dir}(\mathbf{p} \mid \mathbf{1})$. This causes the KL-divergence to reduce to an entropy modulo a constant value that doesn't depend on $\boldsymbol{\phi}$.

Eqn. 14 is identical to the Bayesian loss function in Eqn. 7, 13 and suggests that λ should be set to $1/N^{\mathbf{x}}$ i.e. the inverse of the "pseudo-count" of $\mathbf{x}$ in the training dataset. Intuitively, the entropy regularization is obtained from a singular prior term $\mathrm{Dir}(\mathbf{p} \mid \mathbf{1})$, whereas the cross-entropy loss is obtained from the summation of log-likelihoods over every label observed for $\mathbf{x}$. Therefore, the relative weight of the the two terms is a function of the number of data points $N^{\mathbf{x}}$ in the dataset. The higher the value of $N^{\mathbf{x}}$, the lower is the relative weight of the entropy regularization term.

Since the true pseudo-count is not available, we use the predicted pseudo-count $N^{\boldsymbol{\phi}}(\mathbf{x}) = N_H p_{\boldsymbol{\theta}}(\mathbf{x})$. Our proposed *density-corrected* loss function is:

$$\arg\min_{\boldsymbol{\phi}} \sum_n \Big[-\mathbb{E}_{\mathbf{p}\sim\pi_{\boldsymbol{\phi}}(\mathbf{x})} \log p(y^{(n)} \mid \mathbf{p}) - \underbrace{\frac{1}{N^{\boldsymbol{\phi}}(\mathbf{x})}}_{\text{density-based correction}} \mathbb{H}(\pi_{\boldsymbol{\phi}}(\mathbf{x})) \Big]$$

Charpentier et al. [6] set $\lambda = 10^{-5}$. This implicitly corresponding to $N^{\mathbf{x}} = 10^5$. Similarly, Charpentier et al. [7] perform a grid search for λ in the range $[0, 10^{-5}]$. Importantly, $N^{\mathbf{x}}$ is treated to be uniform across all $\mathbf{x}$ in the training set. In contrast, our proposed coefficient is a function of the density of the individual input $\mathbf{x}$, and cannot be replicated by grid-search (of any resolution) over uniform values.

D. Analyzing the Bayesian loss function with our proposed coefficient

In order to analyze the properties of the Bayesian loss function and motivate our choice of the coefficient $\lambda = \frac{1}{N^{\boldsymbol{\phi}}(\mathbf{x})}$ from a different perspective, we will take the partial derivative (∂) of the Bayesian loss function. We will treat the Bayesian loss function $\mathcal{L}$ in Eqn. 13 as a function $\mathcal{L}(\boldsymbol{\beta}^{\boldsymbol{\phi}}, N^{\boldsymbol{\phi}}, \lambda)$ of three variables: (i) $\boldsymbol{\beta}^{\boldsymbol{\phi}} \in \Delta^C$, (ii) $N^{\boldsymbol{\phi}} \in \mathbb{R}_{\geq 0}$ and (iii) $\lambda \in \mathbb{R}$. Here, λ may potentially be a function of network parameters $\boldsymbol{\phi}$. The partial derivative is taken with respect to a generic quantity and not specialized for now.

$$\begin{aligned}
&\partial\ \mathcal{L}(\boldsymbol{\beta}^{\boldsymbol{\phi}}, N^{\boldsymbol{\phi}}, \lambda) \\
&= \partial \left\{ -\frac{1}{N^{\mathbf{x}}} \sum_{n=1}^{N^{\mathbf{x}}} \mathbb{E}_{\mathbf{p}\sim\pi_{\boldsymbol{\phi}}} \big[\log p(y^{(n)} \mid \mathbf{p}) \big] - \lambda\ \mathbb{H}(\mathrm{Dir}(\mathbf{p} \mid \boldsymbol{\alpha}^{\boldsymbol{\phi}})) \right\} \\
&= -\partial \sum_{c=1}^{C} \frac{N_c^{\mathbf{x}}}{N^{\mathbf{x}}}\ \mathbb{E}_{\mathbf{p}\sim\pi_{\boldsymbol{\phi}}} \big[\log\ p_c \big] - \partial \Big(\lambda\ \mathbb{H}(\mathrm{Dir}(\mathbf{p} \mid \boldsymbol{\alpha}^{\boldsymbol{\phi}})) \Big) \\
&= -\partial \sum_{c=1}^{C} \beta_c^{\mathbf{x}}\ (\psi(1 + N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}) - \psi(C + N^{\boldsymbol{\phi}})) \\
&\quad - \partial \Big(\lambda \Big[\log \mathrm{B}(\mathbf{1} + N^{\boldsymbol{\phi}}\boldsymbol{\beta}^{\boldsymbol{\phi}}) + N^{\boldsymbol{\phi}}\psi(C + N^{\boldsymbol{\phi}} \\
&\qquad - \sum_{c=1}^{C} N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}\ \psi(1 + N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}) \Big] \Big) \\
&= -\partial \sum_{c=1}^{C} \beta_c^{\mathbf{x}}\ \psi(1 + N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}) - \psi(C + N^{\boldsymbol{\phi}})
\end{aligned}$$

$$
\begin{aligned}
&- \partial\Big(\lambda\Big[\sum_{c=1}^{C}\log\Gamma(1+N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}) - \log\Gamma(C+N^{\boldsymbol{\phi}}) \\
&\qquad + N^{\boldsymbol{\phi}}\psi(C+N^{\boldsymbol{\phi}}) - \sum_{c=1}^{C} N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}\ \psi(1+N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\Big]\Big) \\
&= -\sum_{c=1}^{C}\beta_c^{\mathbf{x}}\ \psi'(1+N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\ \partial(N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}) + \psi'(C+N^{\boldsymbol{\phi}})\ \partial N^{\boldsymbol{\phi}} \\
&\quad - \lambda\Bigg[\sum_{c=1}^{C}\cancel{\psi(1+N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\ \partial(N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})} - \cancel{\psi(C+N^{\boldsymbol{\phi}})\ \partial N^{\boldsymbol{\phi}}} \\
&\quad + \cancel{\psi(C+N^{\boldsymbol{\phi}})\ \partial N^{\boldsymbol{\phi}}} + N^{\boldsymbol{\phi}}\psi'(C+N^{\boldsymbol{\phi}})\ \partial N^{\boldsymbol{\phi}} \\
&\quad - \sum_{c=1}^{C}\cancel{\psi(1+N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\ \partial(N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})} \\
&\quad - \sum_{c=1}^{C} N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}\ \psi'(1+N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\ \partial(N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\Bigg] \\
&\quad - (\partial\lambda)\ \mathbb{H}(\mathrm{Dir}(\mathbf{1}+N^{\boldsymbol{\phi}}\boldsymbol{\beta}^{\boldsymbol{\phi}})) \\
&= -\sum_{c=1}^{C}\beta_c^{\mathbf{x}}\ \psi'(1+N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\ \partial(N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}) + \psi'(C+N^{\boldsymbol{\phi}})\ \partial N^{\boldsymbol{\phi}} \\
&\quad - \lambda N^{\boldsymbol{\phi}}\psi'(C+N^{\boldsymbol{\phi}})\ \partial N^{\boldsymbol{\phi}} \\
&\quad + \lambda\sum_{c=1}^{C} N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}\ \psi'(1+N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\partial(N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}) \\
&\quad - (\partial\lambda)\ \mathbb{H}(\mathrm{Dir}(\mathbf{1}+N^{\boldsymbol{\phi}}\boldsymbol{\beta}^{\boldsymbol{\phi}})) \\
&= -\sum_{c=1}^{C}\psi'(1+N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\ \partial(N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\ \Big[\beta_c^{\mathbf{x}} - \lambda\ N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}\Big] \\
&\quad + \psi'(C+N^{\boldsymbol{\phi}})\ \partial(N^{\boldsymbol{\phi}})\ \Big[1-\lambda N^{\boldsymbol{\phi}}\Big] \\
&\quad - (\partial\lambda)\ \mathbb{H}(\mathrm{Dir}(\mathbf{1}+N^{\boldsymbol{\phi}}\boldsymbol{\beta}^{\boldsymbol{\phi}}))
\end{aligned}
$$

where $\Gamma(z)$ is the Gamma function [59], $\psi(z) = \frac{\mathrm{d}}{\mathrm{d}z}\log\Gamma(z)$ is the digamma function [57], and $\mathrm{B}(\mathbf{z})$ is the multivariate beta function [56] where for $\mathbf{z} = (z_1, \ldots, z_k)$, $\mathrm{B}(\mathbf{z}) = \prod_{i=1}^{K}\Gamma(z_i)/\Gamma\big(\sum_{i=1}^{K} z_i\big)$. To summarize, the partial derivative of the Bayesian loss is:

$$
\begin{aligned}
\partial\mathcal{L}(\boldsymbol{\beta}^{\boldsymbol{\phi}}, N^{\boldsymbol{\phi}}, \lambda) = &-\sum_{c=1}^{C}\psi'(1+N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\ \partial(N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\ \Big[\beta_c^{\mathbf{x}} - \lambda\ N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}\Big] \\
&+ \psi'(C+N^{\boldsymbol{\phi}})\ \partial(N^{\boldsymbol{\phi}})\ \Big[1-\lambda N^{\boldsymbol{\phi}}\Big] \\
&- (\partial\lambda)\ \mathbb{H}(\mathrm{Dir}(\mathbf{1}+N^{\boldsymbol{\phi}}\boldsymbol{\beta}^{\boldsymbol{\phi}})) \qquad (15)
\end{aligned}
$$

We will now derive certain desirable properties of $\lambda = \frac{1}{N^{\boldsymbol{\phi}}}$.

Fortunately, the following theorems hold when using the predicted pseudo-count $N^{\boldsymbol{\phi}}(\mathbf{x})$ in λ, even though Eqn. 14 requires the true but unknown pseudo-count $N^{\mathbf{x}}$, and $N^{\boldsymbol{\phi}}(\mathbf{x})$ could be an arbitrarily bad approximation of $N^{\mathbf{x}}$.

Theorem 1: **Necessity**

When $\lambda \neq \frac{1}{N^{\boldsymbol{\phi}}(\mathbf{x})}$, the true posterior may not minimize the Bayesian loss function.

Proof: To see this, let us compute the extremum of the Bayesian loss by computing its partial derivatives with respect to $\beta_c^{\boldsymbol{\phi}}$ i.e. $\frac{\partial}{\partial\beta_c^{\boldsymbol{\phi}}}\mathcal{L}(\boldsymbol{\beta}^{\boldsymbol{\phi}}, N^{\boldsymbol{\phi}}, \lambda)$. Note that since $\boldsymbol{\beta}^{\boldsymbol{\phi}} \in \Delta^C$, it has $C-1$ degrees of freedom. Let us parametrize it by $\{\beta_c^{\boldsymbol{\phi}}\}_{c=1}^{C-1}$ where $\beta_C^{\boldsymbol{\phi}}$ is a dependent variable — $\beta_C^{\boldsymbol{\phi}} = 1 - \sum_{c=1}^{C-1}\beta_c^{\boldsymbol{\phi}}$. This provides the partial derivative with respect to $\beta_c^{\boldsymbol{\phi}}$ as:

$$
\begin{aligned}
\frac{\partial\mathcal{L}(\boldsymbol{\beta}^{\boldsymbol{\phi}}, N^{\boldsymbol{\phi}}, \lambda)}{\partial\beta_c^{\boldsymbol{\phi}}} = -N^{\boldsymbol{\phi}}\Big(&\psi'(1+N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\ \Big[\beta_c^{\mathbf{x}} - \lambda\ N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}\Big] \\
&+ \psi'(1+N^{\boldsymbol{\phi}}\beta_C^{\boldsymbol{\phi}})\ \Big[\beta_C^{\mathbf{x}} - \lambda\ N^{\boldsymbol{\phi}}\beta_C^{\boldsymbol{\phi}}\Big]\Big) \qquad (16)
\end{aligned}
$$

Observe that $\boldsymbol{\beta}^{\boldsymbol{\phi}} = \boldsymbol{\beta}^{\mathbf{x}}/\lambda N^{\boldsymbol{\phi}}$ is an extremum. Therefore, if $\lambda \neq 1/N^{\boldsymbol{\phi}}$, the minimization of the Bayesian loss might not result in the true posterior $\boldsymbol{\beta}^{\boldsymbol{\phi}} = \boldsymbol{\beta}^{\mathbf{x}}$. □

Theorem 2: **Extremity**

When $\lambda = \frac{1}{N^{\boldsymbol{\phi}}(\mathbf{x})}$, the true posterior is an extremum of the Bayesian loss function.

Proof: Let us first simplify Eqn. 15 by setting $\lambda = \frac{1}{N^{\boldsymbol{\phi}}}$.

$$
\begin{aligned}
\partial\mathcal{L}(\boldsymbol{\beta}^{\boldsymbol{\phi}}, N^{\boldsymbol{\phi}}) &= -\sum_{c=1}^{C}\psi'(1+N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\ \partial(N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\Big[\beta_c^{\mathbf{x}} - \beta_c^{\boldsymbol{\phi}}\Big] \\
&\quad - \Big(\frac{\partial}{\partial N^{\boldsymbol{\phi}}}\frac{1}{N^{\boldsymbol{\phi}}}\Big)(\partial N^{\boldsymbol{\phi}})\ \mathbb{H}(\mathrm{Dir}(\mathbf{1}+N^{\boldsymbol{\phi}}\boldsymbol{\beta}^{\boldsymbol{\phi}})) \\
&= -\sum_{c=1}^{C}\psi'(1+N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\ \partial(N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\Big[\beta_c^{\mathbf{x}} - \beta_c^{\boldsymbol{\phi}}\Big] \\
&\quad + (\partial N^{\boldsymbol{\phi}})\frac{1}{(N^{\boldsymbol{\phi}})^2}\ \mathbb{H}(\mathrm{Dir}(\mathbf{1}+N^{\boldsymbol{\phi}}\boldsymbol{\beta}^{\boldsymbol{\phi}})) \qquad (17)
\end{aligned}
$$

Let us also simplify Eqn. 16 by setting $\lambda = \frac{1}{N^{\boldsymbol{\phi}}}$.

$$
\begin{aligned}
\frac{\partial\mathcal{L}(\boldsymbol{\beta}^{\boldsymbol{\phi}}, N^{\boldsymbol{\phi}})}{\partial\beta_c^{\boldsymbol{\phi}}} = -N^{\boldsymbol{\phi}}\Big(&\psi'(1+N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\ \Big[\beta_c^{\mathbf{x}} - \beta_c^{\boldsymbol{\phi}}\Big] \\
&- \psi'(1+N^{\boldsymbol{\phi}}\beta_C^{\boldsymbol{\phi}})\ \Big[\beta_C^{\mathbf{x}} - \beta_C^{\boldsymbol{\phi}}\Big]\Big) \qquad (18)
\end{aligned}
$$

Eqn 18 implies than when $\boldsymbol{\beta}^{\boldsymbol{\phi}} = \boldsymbol{\beta}^{\mathbf{x}}$, then $\partial\mathcal{L}/\partial\beta_c^{\boldsymbol{\phi}} = 0$. Therefore, $\boldsymbol{\beta}^{\boldsymbol{\phi}} = \boldsymbol{\beta}^{\mathbf{x}}$ is an extremum of the Bayesian loss function. □

However, this does not preclude the existence of other extrema where $\boldsymbol{\beta}^{\boldsymbol{\phi}} \neq \boldsymbol{\beta}^{\mathbf{x}}$ that minimize the Bayesian loss function. Fortunately, the next property rules out this possibility.

Theorem 3: **Uniqueness**

When $\lambda = \frac{1}{N^{\boldsymbol{\phi}}(\mathbf{x})}$, the true posterior is the unique extremum of the Bayesian loss function.

Proof: In general, a system of $C-1$ non-linear equations in $C-1$ variables given by Eqn. 16 can have multiple solutions, since the digamma function ψ is non-linear. However, we show that in the special case of $\lambda = 1/N^{\boldsymbol{\phi}}$, the solution is unique. To prove that the extremum is unique, we will use

the fact that the derivative of the digamma function is always positive [57] : $\psi'(z) > 0,\ \forall z > 0$. Setting Eqn. 18 to equal zero, we get

$$\forall c \in \{1, \ldots, C\} \ : \ \psi'(1 + N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}) \left[\beta_c^{\mathbf{x}} - \beta_c^{\boldsymbol{\phi}}\right] = \gamma \quad (19)$$

for some constant γ that depends on $N^{\boldsymbol{\phi}}$ and $\boldsymbol{\beta}^{\boldsymbol{\phi}}$. We will now prove that $\gamma = 0$.

First, assume that $\gamma > 0$. Then, since ψ' is always positive, $\psi'(1 + N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}) \left[\beta_c^{\mathbf{x}} - \beta_c^{\boldsymbol{\phi}}\right] > 0 \implies \beta_c^{\mathbf{x}} - \beta_c^{\boldsymbol{\phi}} > 0 \ ; \ \forall c$. Summing the C equations given by Eqn. 19, since $\sum_{c=1}^{C} \beta_c^{\mathbf{x}} = \sum_{c=1}^{C} \beta_c^{\boldsymbol{\phi}} = 1$, we get $0 > 0$, which is a contradiction. Similarly, assuming $\gamma < 0$ also leads to a contradiction. This implies that $\gamma = 0$ which further implies that $\psi'(1 + N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}) \left[\beta_c^{\mathbf{x}} - \beta_c^{\boldsymbol{\phi}}\right] = 0 \implies \beta_c^{\mathbf{x}} = \beta_c^{\boldsymbol{\phi}} \ ; \ \forall c$. This means that $\boldsymbol{\beta}^{\boldsymbol{\phi}} = \boldsymbol{\beta}^{\mathbf{x}}$ is the unique extremum. □

Theorem 4: **Minimum**

When $\lambda = \frac{1}{N^{\boldsymbol{\phi}}(\mathbf{x})}$, the unique extremum of the Bayesian loss function is a minimum.

Proof: We compute the second derivative $\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}^{\boldsymbol{\phi}}, N^{\boldsymbol{\phi}})}{\partial\ (\beta_c^{\boldsymbol{\phi}})^2}$ by differentiating Eqn. 18 with respect to $\beta_c^{\boldsymbol{\phi}}$:

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}^{\boldsymbol{\phi}}, N^{\boldsymbol{\phi}})^2}{\partial {\beta_c^{\boldsymbol{\phi}}}^2}$$

$$= -(N^{\boldsymbol{\phi}})^2 \underbrace{\left(\psi''(1 + N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}) \left[\beta_c^{\mathbf{x}} - \beta_c^{\boldsymbol{\phi}}\right] + \psi''(1 + N^{\boldsymbol{\phi}}\beta_C^{\boldsymbol{\phi}}) \left[\beta_C^{\mathbf{x}} - \beta_C^{\boldsymbol{\phi}}\right]\right)}_{= \ 0 \ \text{at} \ \boldsymbol{\beta}^{\boldsymbol{\phi}} = \boldsymbol{\beta}^{\mathbf{x}}}$$

$$+ N^{\boldsymbol{\phi}} \underbrace{\left(\psi'(1 + N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}) + \psi'(1 + N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}})\right)}_{> \ 0} > 0$$

The first term is zero at $\boldsymbol{\beta}^{\boldsymbol{\phi}} = \boldsymbol{\beta}^{\mathbf{x}}$, whereas the second term is always positive since $\psi'(z) > 0,\ \forall z > 0$ [57]. Since the second derivative is positive at the unique extremum, the extremum is a minimum. □

Theorem 5: **Density maximization**

When $\lambda = \frac{1}{N^{\boldsymbol{\phi}}(\mathbf{x})}$, the Bayesian loss function is minimized as $N^{\boldsymbol{\phi}}(\mathbf{x}) \to +\infty$. This trains the normalizing flow to maximize predicted density $p_{\boldsymbol{\theta}}(\mathbf{x})$ on the training data.

Proof: Assuming that $\lambda = 1/N^{\boldsymbol{\phi}}$ and $\frac{\partial}{\partial N^{\boldsymbol{\phi}}} \beta_c^{\boldsymbol{\phi}} = 0$, from Eqn. 17 we get:

$$0 = \frac{\partial \mathcal{L}(\boldsymbol{\beta}^{\boldsymbol{\phi}}, N^{\boldsymbol{\phi}})}{\partial N^{\boldsymbol{\phi}}}$$

$$= -\sum_{c=1}^{C} \psi'(1 + N^{\boldsymbol{\phi}}\beta_c^{\boldsymbol{\phi}}) \ \beta_c^{\boldsymbol{\phi}} \left[\beta_c^{\mathbf{x}} - \beta_c^{\boldsymbol{\phi}}\right] + \frac{1}{(N^{\boldsymbol{\phi}})^2} \ \mathbb{H}(\text{Dir}(\mathbf{1} + N^{\boldsymbol{\phi}}\boldsymbol{\beta}^{\boldsymbol{\phi}}))$$

$$= \frac{1}{(N^{\boldsymbol{\phi}})^2} \ \mathbb{H}(\text{Dir}(\mathbf{1} + N^{\boldsymbol{\phi}}\boldsymbol{\beta}^{\boldsymbol{\phi}})) \quad (\boldsymbol{\beta}^{\boldsymbol{\phi}} = \boldsymbol{\beta}^{\mathbf{x}} \text{ at the unique optimum})$$

At the optimum, the entropy of the Dirichlet $\mathbb{H}(\text{Dir}(\mathbf{p} \mid \mathbf{1} + N^{\boldsymbol{\phi}}\boldsymbol{\beta}^{\boldsymbol{\phi}}))$ must be zero. This can only happen when the concentration parameter of the Dirichlet i.e. $N^{\boldsymbol{\phi}}$ increases to $+\infty$. Therefore, when $\lambda = \frac{1}{N^{\boldsymbol{\phi}}}$, the Bayesian loss function tries to increase the density of the normalizing flow $N^{\boldsymbol{\phi}}(\mathbf{x}) = N_H \ p_{\boldsymbol{\theta}}(\mathbf{x})$ on in-distribution examples in the training set to infinity. This is similar to the conventional log-likelihood training objective for normalizing flows, which also maximizes the probability density $p_{\boldsymbol{\theta}}(\mathbf{x})$ predicted by the normalizing flow on the training set. Therefore, the Bayesian loss trains both the classifier $\boldsymbol{\beta}^{\boldsymbol{\phi}}$ and the normalizing flow $p_{\boldsymbol{\phi}}$ provided $\lambda = \frac{1}{N^{\boldsymbol{\phi}}}$. □

Note that theorems 1-5 hold when using the predicted pseudo-count $N^{\boldsymbol{\phi}}(\mathbf{x})$ for λ, even though Eqn. 14 requires the true but unknown pseudo-count $N^{\mathbf{x}}$, and $N^{\boldsymbol{\phi}}(\mathbf{x})$ could be an arbitrarily bad approximation of $N^{\mathbf{x}}$.