
Being Bayesian about Categorical Probability

Taejong Joo¹ Uijung Chung¹ Min-Gwan Seo¹

Abstract

Neural networks utilize the softmax as a building block in classification tasks, which contains an overconfidence problem and lacks an uncertainty representation ability. As a Bayesian alternative to the softmax, we consider a random variable of a categorical probability over class labels. In this framework, the prior distribution explicitly models the presumed noise inherent in the observed label, which provides consistent gains in generalization performance in multiple challenging tasks. The proposed method inherits advantages of Bayesian approaches that achieve better uncertainty estimation and model calibration. Our method can be implemented as a plug-and-play loss function with negligible computational overhead compared to the softmax with the cross-entropy loss function.

1. Introduction

Softmax (Bridle, 1990) is the de facto standard for post processing of logits of neural networks (NNs) for classification. When combined with the maximum likelihood objective, it enables efficient gradient computation with respect to logits and has achieved state-of-the-art performances on many benchmark datasets. However, softmax lacks the ability to represent the uncertainty of predictions (Blundell et al., 2015; Gal & Ghahramani, 2016) and has poorly calibrated behavior (Guo et al., 2017). For instance, the NN with softmax can easily be fooled to confidently produce wrong outputs; when rotating digit 3, it will predict it as the digit 8 or 4 with high confidence (Louizos & Welling, 2017). Another concern of the softmax is its confident predictive behavior makes NNs to be subject to overfitting (Xie et al., 2016; Pereyra et al., 2017). This issue raises the need for effective regularization techniques for improving generalization performance.

¹ESTsoft, Republic of Korea. Correspondence to: Taejong Joo <tjoo@estsoft.com>.

Bayesian NNs (BNNs; MacKay, 1992) can address the aforementioned issues of softmax. BNNs provide quantifiable measures of uncertainty such as predictive entropy and mutual information (Gal, 2016) and enable automatic embodiment of Occam’s razor (MacKay, 1995). However, some practical obstacles have impeded the wide adoption of BNNs. First, the intractable posterior inference in BNNs demands approximate methods such as variational inference (VI; Graves, 2011; Blundell et al., 2015) and Monte Carlo (MC) dropout (Gal & Ghahramani, 2016). Even with such novel approximation methods, concerns arise regarding both the degree of approximation and the computational expensive posterior inference (Wu et al., 2019a; Osawa et al., 2019). In addition, under extreme non-linearity between parameters and outputs in the NNs, determining a meaningful weight prior distribution is challenging (Sun et al., 2019). Last but not least, BNNs often require considerable modifications to existing baselines, or they result in performance degradation (Lakshminarayanan et al., 2017).

In this paper, we apply the Bayesian principle to construct the target distribution for learning classifiers. Specifically, we regard a categorical probability as a random variable, and construct the target distribution over the categorical probability by means of the Bayesian inference, which is approximated by NNs. The resulting target distribution can be thought of as being regularized via the prior belief whose impact is controlled by the number of observations. By considering only the random variable of categorical probability, the Bayesian principle can be efficiently adopted to existing deep learning building blocks without huge modifications. Our extensive experiments show effectiveness of being Bayesian about the categorical probability in improving generalization performances, uncertainty estimation, and calibration property.

Our contributions can be summarized as follows: 1) we show the importance of considering categorical probability as a random variable instead of being determined by the label; 2) we provide experimental results showing the usefulness of the Bayesian principle in improving generalization performance of large models on standard benchmark datasets, e.g., ResNext-101 on ImageNet; 3) we enable NNs to inherit the advantages of the Bayesian methods in better uncertainty representation and well-calibrated behavior with a negligible increase in computational complexity.

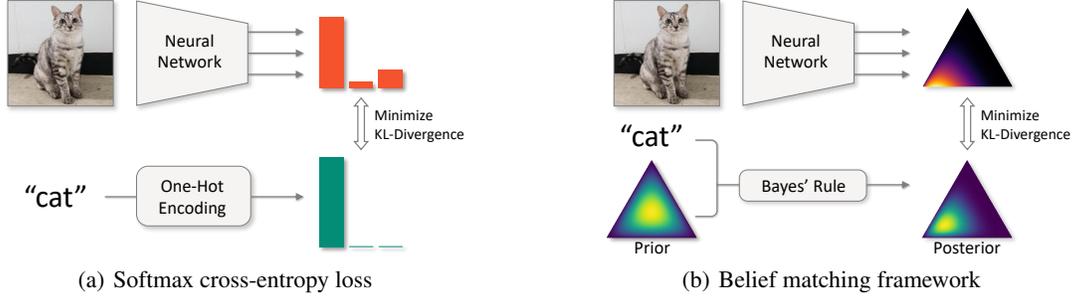


Figure 1. Illustration of the difference between softmax cross-entropy loss and belief matching framework when each image is unique in the training set. In softmax cross-entropy loss, the label “cat” is directly transformed into the target categorical distribution. In belief matching framework, the label “cat” is combined with the prior Dirichlet distribution over the categorical probability. Then, the Bayes’ rule updates the *belief* about categorical probability, which produces the target distribution.

2. Preliminary

This paper focuses on classification problems in which, given i.i.d. training samples $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N \in (\mathcal{X} \times \mathcal{Y})^N$, we construct a classifier $\mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y}$. Here, \mathcal{X} is an input space and $\mathcal{Y} = \{1, \dots, K\}$ is a set of labels. We denote \mathbf{x} and y as random variables whose unknown probability distributions generate inputs and labels, respectively. Also, we let $\tilde{\mathbf{y}}$ be a one-hot representation of y .

Let $f^{\mathbf{W}} : \mathcal{X} \rightarrow \mathcal{X}'$ be a NN with parameters \mathbf{W} where $\mathcal{X}' = \mathbb{R}^K$ is a logit space. In this paper, we assume $\arg \max_j f_j^{\mathbf{W}} \subseteq \mathcal{F}$ is the classification model where $f_j^{\mathbf{W}}$ denotes the j -th output basis of $f^{\mathbf{W}}$, and we concentrate on the problem of learning \mathbf{W} . Given $((\mathbf{x}, y) \in \mathcal{D}, f^{\mathbf{W}})$, a standard minimization loss function is the softmax cross-entropy loss, which applies the softmax to logit and then computes the cross-entropy between a one-hot encoded label and a softmax output (Figure 1(a)). Specifically, the softmax, denoted by $\phi : \mathcal{X}' \rightarrow \Delta^{K-1}$, transforms a logit $f^{\mathbf{W}}(\mathbf{x})$ into a normalized exponential form:

$$\phi_k(f^{\mathbf{W}}(\mathbf{x})) = \frac{\exp(f_k^{\mathbf{W}}(\mathbf{x}))}{\sum_j \exp(f_j^{\mathbf{W}}(\mathbf{x}))} \quad (1)$$

, and then the cross-entropy loss can be computed by $l_{CE}(\tilde{\mathbf{y}}, \phi(f^{\mathbf{W}}(\mathbf{x}))) = -\sum_k \tilde{\mathbf{y}}_k \log \phi_k(f^{\mathbf{W}}(\mathbf{x}))$. Here, note that the softmax output can be viewed as a parameter of the categorical distribution, which can be denoted by $\mathcal{P}^C(\phi(f^{\mathbf{W}}(\mathbf{x})))$.

We can formulate the minimization of the softmax cross-entropy loss over \mathcal{D} into a collection of distribution matching problems. To this end, let $c^{\mathcal{D}}(\mathbf{x})$ be a vector-valued function that counts label frequency at $\mathbf{x} \in \mathcal{X}$ in \mathcal{D} , which is defined as:

$$c^{\mathcal{D}}(\mathbf{x}) = \sum_{(\mathbf{x}', y') \in \mathcal{D}} \tilde{\mathbf{y}}' \mathbb{1}_{\{\mathbf{x}\}}(\mathbf{x}') \quad (2)$$

where $\mathbb{1}_{\mathcal{A}}(x)$ is an indicator function that takes 1 when

$x \in \mathcal{A}$ and 0 otherwise. Then, the empirical risk on \mathcal{D} can be expressed as follows:

$$\begin{aligned} \hat{\mathcal{L}}_{\mathcal{D}}(\mathbf{W}) &= -\frac{1}{N} \sum_{i=1}^N \log \phi_{y^{(i)}}(f^{\mathbf{W}}(\mathbf{x}^{(i)})) \\ &= \sum_{\mathbf{x} \in G(\mathcal{D})} \frac{\sum_i c_i^{\mathcal{D}}(\mathbf{x})}{N} l_{\mathbf{x}}^{\mathcal{D}}(\mathbf{W}) + C \quad (3) \end{aligned}$$

where $G(\mathcal{D})$ is a set of unique values in \mathcal{D} , e.g., $G(\{1, 2, 2\}) = \{1, 2\}$, and C is a constant with respect to \mathbf{W} ; $l_{\mathbf{x}}^{\mathcal{D}}(\mathbf{W})$ measures the KL divergence between the empirical target distribution and the categorical distribution modeled by the NN at location \mathbf{x} , which is given by:

$$l_{\mathbf{x}}^{\mathcal{D}}(\mathbf{W}) = KL \left(\mathcal{P}^C \left(\frac{c^{\mathcal{D}}(\mathbf{x})}{\sum_i c_i^{\mathcal{D}}(\mathbf{x})} \right) \parallel \mathcal{P}^C(\phi(f^{\mathbf{W}}(\mathbf{x}))) \right) \quad (4)$$

Therefore, the normalized value of $c^{\mathcal{D}}(\mathbf{x})$ becomes the *estimator of a categorical probability* of the target distribution at location \mathbf{x} . However, directly approximating this target distribution can be problematic. This is because the estimator uses single or very few samples since most of the inputs are unique or very rare in the training set.

One simple heuristic to handle this problem is label smoothing (Szegedy et al., 2016) that constructs a regularized target estimator, in which a one-hot encoded label $\tilde{\mathbf{y}}$ is relaxed by $(1 - \lambda)\tilde{\mathbf{y}} + \frac{\lambda}{K}\mathbf{1}$ with hyperparameter λ . Under the smoothing operation, the target estimator is regularized by a mixture of the empirical counts and the parameter of the discrete uniform distribution \mathcal{P}^U such that $(1 - \lambda)\mathcal{P}^C \left(\frac{c^{\mathcal{D}}(\mathbf{x})}{\sum_i c_i^{\mathcal{D}}(\mathbf{x})} \right) + \lambda\mathcal{P}^U$. One concern is that the mixing coefficient is constant with respect to the number of observations, which can possibly prevent the exploitation of the empirical counting information when it is needed.

Another more principled approach is BNNs, which prevents full exploitation of the noisy estimation by balancing the distance to the target distribution with model

complexity and maintaining the weight ensemble instead of choosing a single best configuration. Specifically, in BNNs with the Gaussian weight prior $\mathcal{N}(\mathbf{0}, \tau^{-1}\mathbf{I})$, the *score* of configuration \mathbf{W} is measured by the posterior density $p_{\mathbf{W}}(\mathbf{W}|\mathcal{D}) \propto p(\mathcal{D}|\mathbf{W})p_{\mathbf{W}}(\mathbf{W})$ where we have $\log p_{\mathbf{W}}(\mathbf{W}) \propto -\tau \|\mathbf{W}\|_2^2$. Therefore, the complexity penalty term induced by the prior prevents the softmax output from exactly matching a one-hot encoded target. In modern deep NNs, however, $\|\mathbf{W}\|_2^2$ may be poor proxy for the model complexity due to extreme non-linear relationship between weights and outputs (Hafner et al., 2018; Sun et al., 2019) as well as weight-scaling invariant property of batch normalization (Ioffe & Szegedy, 2015). This issue may result in poorly regularized predictions, i.e., cannot prevent NNs from the full exploitation of the information contained in the noisy target estimator.

3. Method

3.1. Constructing Target Distribution

We propose a Bayesian approach to construct the target distribution for classification, called a belief matching framework (BM; Figure 1(b)), in which the categorical probability about a label is regarded as a random variable \mathbf{z} . Specifically, we express the likelihood of \mathbf{z} (given \mathbf{x}) about the label y as a categorical distribution $p_{y|\mathbf{x},\mathbf{z}} = \mathcal{P}^C(\mathbf{z}|\mathbf{x})^1$. Then, specification of the prior distribution over $\mathbf{z}|\mathbf{x}$ automatically determines the target distribution by means of the Bayesian inference: $p_{\mathbf{z}|\mathbf{x},y}(\mathbf{z}) \propto p_{y|\mathbf{z},\mathbf{x}}(y)p_{\mathbf{z}|\mathbf{x}}(\mathbf{z})$.

We consider a conjugate prior for simplicity, i.e., the Dirichlet distribution. A random variable \mathbf{z} (given \mathbf{x}) following the Dirichlet distribution with concentration parameter vector β , denoted by $\mathcal{P}^D(\beta)$, has the following density:

$$p_{\mathbf{z}|\mathbf{x}}(\mathbf{z}) = \frac{\Gamma(\beta_0)}{\prod_j \Gamma(\beta_j)} \prod_{k=1}^K z_k^{\beta_k - 1} \quad (5)$$

where $\Gamma(\cdot)$ is the gamma function, $\sum_i z_i = 1$ meaning that \mathbf{z} belongs to the $K - 1$ simplex Δ^{K-1} , $\beta_i > 0$, $\forall i$, and $\beta_0 = \sum_i \beta_i$. Here, we have that the mean of $\mathbf{z}|\mathbf{x}$ is β/β_0 and β_0 controls the sharpness of the density such that more mass centered around the mean as β_0 becomes larger.

By the characteristics of the conjugate family, we have the following posterior distribution given \mathcal{D} :

$$p_{\mathbf{z}|\mathbf{x},y} = \mathcal{P}^D(\beta + c^{\mathcal{D}}(\mathbf{x})) \quad (6)$$

where the target posterior mean is explicitly *smoothed* by the prior belief, and the smoothing operation is performed by the principled way of applying Bayes' rule. Specifically, the posterior mean is given by $\frac{1}{\beta_0 + \sum_i c_i^{\mathcal{D}}(\mathbf{x})}(\beta + c^{\mathcal{D}}(\mathbf{x}))$, in

¹In this paper, $p_{\mathbf{x}} = \mathcal{P}(\theta)$ is read as a random variable \mathbf{x} follows a probability distribution \mathcal{P} with parameter θ .

which the prior distribution acts as adding pseudo counts. We note that the relative strength between the prior belief and the empirical count information becomes adaptive with respect to each data point.

3.2. Representing Approximate Distribution

Now, we specify the approximate posterior distribution modeled by the NNs, which aims to approximate $p_{\mathbf{z}|\mathbf{x},y}$. In this paper, we model the approximate posterior as the Dirichlet distribution. To this end, we use an exponential function to transform logits to the concentration parameter of \mathcal{P}^D , and we let $\alpha^{\mathbf{W}} = \exp \circ f^{\mathbf{W}}$. Then, the NN represents the density over Δ^{K-1} as follows:

$$q_{\mathbf{z}|\mathbf{x}}^{\mathbf{W}}(\mathbf{z}) = \frac{\Gamma(\alpha_0^{\mathbf{W}}(\mathbf{x}))}{\prod_j \Gamma(\alpha_j^{\mathbf{W}}(\mathbf{x}))} \prod_{k=1}^K z_k^{\alpha_k^{\mathbf{W}}(\mathbf{x}) - 1} \quad (7)$$

where $\alpha_0^{\mathbf{W}}(\mathbf{x}) = \sum_i \alpha_i^{\mathbf{W}}(\mathbf{x})$.

From equation 7, we can see that outputs under BM encode much more information compared to those under the softmax. Specifically, it can be easily shown that the approximate posterior mean corresponds to the softmax. In this regard, BM enables neural networks to represent more rich information in their outputs, i.e., the density over Δ^{K-1} itself not just a single point on it such as the mean. This capability allows capturing more diverse characteristics of predictions at different locations, such as how much concentrate its density around the center mass point, which can be extremely helpful in many applications. For instance, BM gives a more sophisticated measure of the difference between predictions of two neural networks, which can benefit the consistency-based loss for semi-supervised learning as we will show in section 5.4. Besides, BM represents a more sophisticated measure of predictive uncertainty based on the density over simplex, such as mutual information.

From the perspective of learning the target distribution, BM can be considered as a generalization of softmax in terms of changing the moment matching problem to the distribution matching problem in $\mathcal{P}(\Delta^{K-1})$. To understand the distribution matching objective in BM, we reformulate equation 7 as follows:

$$\begin{aligned} q_{\mathbf{z}|\mathbf{x}}^{\mathbf{W}}(\mathbf{z}) &\propto \exp \left(\sum_k \alpha_k^{\mathbf{W}}(\mathbf{x}) \log z_k - \sum_k \log z_k \right) \\ &\propto \exp \left(-l_{CE}(\phi(f^{\mathbf{W}}(\mathbf{x})), \mathbf{z}) + \frac{KL(\mathcal{P}^U \parallel \mathcal{P}^C(\mathbf{z}))}{\alpha_0^{\mathbf{W}}(\mathbf{x})/K} \right) \end{aligned} \quad (8)$$

In the limit of $q_{\mathbf{z}|\mathbf{x}}^{\mathbf{W}} \rightarrow p_{\mathbf{z}|\mathbf{x},y}$, mean of the target posterior (equation 6) becomes a virtual label, for which individual \mathbf{z} ought to match; the penalty for ambiguous configuration \mathbf{z} is determined by the number of observations. Therefore, the

distribution matching in BM can be thought of as *learning to score a categorical probability* based on closeness to the target posterior mean, in which exploitation of the closeness information is automatically controlled by the data.

3.3. Distribution Matching

We have defined the target distribution $p_{\mathbf{z}|\mathbf{x},y}$ and the approximate distribution modeled by the neural network $q_{\mathbf{z}|\mathbf{x}}^{\mathbf{W}}$. We now present a solution to the distribution matching problem with maximizing the evidence lower bound (ELBO), defined by $l_{EB}(y, \alpha^{\mathbf{W}}(\mathbf{x})) = \mathbb{E}_{q_{\mathbf{z}|\mathbf{x}}^{\mathbf{W}}}[\log p(y|\mathbf{x}, \mathbf{z})] - KL(q_{\mathbf{z}|\mathbf{x}}^{\mathbf{W}} \parallel p_{\mathbf{z}|\mathbf{x}})$. Using the ELBO can be motivated by the following equality (Jordan et al., 1999):

$$\begin{aligned} \log p(y|\mathbf{x}) &= \int q_{\mathbf{z}|\mathbf{x}}^{\mathbf{W}}(z) \log \left(\frac{p(y, z|\mathbf{x})}{p(z|\mathbf{x}, y)} \right) dz \\ &= l_{EB}(y, \alpha^{\mathbf{W}}(\mathbf{x})) - KL(q_{\mathbf{z}|\mathbf{x}}^{\mathbf{W}} \parallel p_{\mathbf{z}|\mathbf{x},y}) \end{aligned} \quad (9)$$

where we can see that maximizing $l_{EB}(y, \alpha^{\mathbf{W}}(\mathbf{x}))$ corresponds to minimizing $KL(q_{\mathbf{z}|\mathbf{x}}^{\mathbf{W}} \parallel p_{\mathbf{z}|\mathbf{x},y})$, i.e., *matching the approximate distribution to the target distribution*, because the KL-divergence is non-negative and $\log p(y|\mathbf{x})$ is a constant with respect to \mathbf{W} . Here, each term in the ELBO can be analytically computed by:

$$\begin{aligned} \mathbb{E}_{q_{\mathbf{z}|\mathbf{x}}}[\log p(y|\mathbf{x}, \mathbf{z})] \\ = \mathbb{E}_{q_{\mathbf{z}|\mathbf{x}}}[\log z_y] = \psi(\alpha_y^{\mathbf{W}}(\mathbf{x})) - \psi(\alpha_0^{\mathbf{W}}(\mathbf{x})) \end{aligned} \quad (10)$$

where $\psi(\cdot)$ is the digamma function (the logarithmic derivative of $\Gamma(\cdot)$), and

$$\begin{aligned} KL(q_{\mathbf{z}|\mathbf{x}}^{\mathbf{W}} \parallel p_{\mathbf{z}|\mathbf{x}}) &= \log \frac{\Gamma(\alpha_0^{\mathbf{W}}(\mathbf{x})) \prod_k \Gamma(\beta_k)}{\prod_k \Gamma(\alpha_k^{\mathbf{W}}(\mathbf{x})) \Gamma(\beta_0)} \\ &+ \sum_k (\alpha_k^{\mathbf{W}}(\mathbf{x}) - \beta_k) (\psi(\alpha_k^{\mathbf{W}}(\mathbf{x})) - \psi(\alpha_0^{\mathbf{W}}(\mathbf{x}))) \end{aligned} \quad (11)$$

where $p_{\mathbf{z}|\mathbf{x}}$ is assumed to be an input independent conjugate prior for simplicity; that is, $p_{\mathbf{z}|\mathbf{x}} = \mathcal{P}^D(\beta)$. With this analytical solution, we maximize the ELBO with mini-batch approximation, which gives the following loss function: $\mathcal{L}(\mathbf{W}) = \mathbb{E}_{\mathbf{x},y}[l_{EB}(y, \alpha^{\mathbf{W}}(\mathbf{x}))] \approx \frac{1}{m} \sum_{i=1}^m l_{EB}(y^{(i)}, \alpha^{\mathbf{W}}(\mathbf{x}^{(i)}))$. We note that computations of the ELBO and its gradient have a complexity of $\mathcal{O}(K)$ per sample, which is equal to those of softmax. This means that BM can preserve the scalability and the efficiency of the existing baseline. We also note that the analytical solution of the ELBO under BM allows to implement the distribution matching loss as a *plug-and-play loss function* applied to the logit directly.

3.4. On Prior Distributions

The success of the Bayesian approach largely depends on how we specify the prior distribution due to its impact on

the resulting posterior distribution. For example, the target posterior mean in equation 6 becomes the counting estimator as $\beta_0 \rightarrow 0$. On the contrary, as β_0 becomes higher, the effect of empirical counting information is weakened, and eventually disappeared in the limit of $\beta_0 \rightarrow \infty$. Therefore, considering that most of the inputs are unique in \mathcal{D} , choosing small β_0 is appropriate for prevents the resulting posterior distribution from being dominated by the prior².

However, a prior distribution with small β_0 implicitly makes $\alpha_0^{\mathbf{W}}(\mathbf{x})$ small, which poses significant challenges on the gradient-based optimization. This is because the gradient of the ELBO is notoriously large in the small-value regimes of $\alpha_0^{\mathbf{W}}(\mathbf{x})$, e.g., $\psi'(0.01) > 10000$. In addition, our various building blocks including normalization (Ioffe & Szegedy, 2015), initialization (He et al., 2015), and architecture (He et al., 2016a) are implicitly or explicitly designed to make $\mathbb{E}[f^{\mathbf{W}}(\mathbf{x})] \approx \mathbf{0}$; that is, $\mathbb{E}[\alpha^{\mathbf{W}}(\mathbf{x})] \approx \mathbf{1}$. Therefore, making $\alpha_0^{\mathbf{W}}(\mathbf{x})$ small can be wasteful or requires huge modifications to the existing building blocks. Also, $\mathbb{E}[\alpha^{\mathbf{W}}(\mathbf{x})] \approx \mathbf{1}$ is encouraged in a sense of natural gradient (Amari, 1998), which improves the conditioning of Fisher information matrix (Schraudolph, 1998; LeCun et al., 1998; Raiko et al., 2012; Wiesler et al., 2014).

In order to resolve the gradient-based optimization challenge in learning the posterior distribution while preventing dominance of the prior distribution, we set $\beta = \mathbf{1}$ for the prior distribution and then multiply λ to the KL divergence term in the ELBO: $l_{EB}^\lambda(y, \alpha^{\mathbf{W}}(\mathbf{x})) = \mathbb{E}_{q_{\mathbf{z}|\mathbf{x}}}[\log p(y|\mathbf{x}, \mathbf{z})] - \lambda KL(q_{\mathbf{z}|\mathbf{x}}^{\mathbf{W}} \parallel \mathcal{P}^D(\mathbf{1}))$. This trick significantly stabilizes the optimization process, while making a local optimal point remains unchanged. To see this, we can compare the gradients of the ELBO and the lambda multiplied ELBO:

$$\begin{aligned} \frac{\partial l_{EB}(y, \alpha^{\mathbf{W}}(\mathbf{x}))}{\partial \alpha_k^{\mathbf{W}}(\mathbf{x})} &= (\tilde{y}_k - (\alpha_k^{\mathbf{W}}(\mathbf{x}) - \beta_k)) \psi'(\alpha_k^{\mathbf{W}}(\mathbf{x})) \\ &- (1 - (\alpha_0^{\mathbf{W}}(\mathbf{x}) - \beta_0)) \psi'(\alpha_0^{\mathbf{W}}(\mathbf{x})) \end{aligned} \quad (12)$$

$$\begin{aligned} \frac{\partial l_{EB}^\lambda(y, \alpha^{\mathbf{W}}(\mathbf{x}))}{\partial \alpha_k^{\mathbf{W}}(\mathbf{x})} &= (\tilde{y}_k - (\tilde{\alpha}_k^{\mathbf{W}}(\mathbf{x}) - \lambda)) \frac{\psi'(\tilde{\alpha}_k^{\mathbf{W}}(\mathbf{x}))}{\psi'(\tilde{\alpha}_0^{\mathbf{W}}(\mathbf{x}))} \\ &- (1 - (\tilde{\alpha}_0^{\mathbf{W}}(\mathbf{x}) - \lambda K)) \end{aligned} \quad (13)$$

where $\tilde{\alpha}_k^{\mathbf{W}}(\mathbf{x}) = \lambda \alpha_k^{\mathbf{W}}(\mathbf{x})$. Here, we can see that a local optimal in equation 12 is achieved when $\alpha^{\mathbf{W}}(\mathbf{x}) = \beta + \tilde{\mathbf{y}}$ and a local optima for equation 13 is $\alpha^{\mathbf{W}}(\mathbf{x}) = 1 + \frac{1}{\lambda} \tilde{\mathbf{y}}$. Therefore, a ratio between $\alpha_i^{\mathbf{W}}(\mathbf{x})$ and $\alpha_j^{\mathbf{W}}(\mathbf{x})$ equal to those of a local optimal point in equation 12 for every pair of i and j . In this regard, searching for λ with $l_{EB}^\lambda(y, \alpha^{\mathbf{W}}(\mathbf{x}))$ and then multiplying λ after training corresponds to the process of searching for the prior distribution's parameter β with $l_{EB}(y, \alpha^{\mathbf{W}}(\mathbf{x}))$.

²In an ideal fully Bayesian treatment, β can be modeled hierarchically, and we left this as future research.

4. Related Work

BNNs are the dominant approach for applying Bayesian principles in neural networks. Because BNNs require the intractable posterior inference, many posterior approximation schemes have been developed to reduce the approximation gap and improve scalability (e.g., VI (Graves, 2011; Blundell et al., 2015; Wu et al., 2019a) and stochastic gradient Markov Chain Monte Carlo (Welling & Teh, 2011; Ma et al., 2015; Gong et al., 2019)). However, even with these novel approximation techniques, BNNs are not scalable to state-of-the-art architectures in large-scale datasets or they often reduce the generalization performance in practice, which impedes the wide adoption of BNNs despite their numerous potential benefits.

Other approaches avoid explicit modeling of the weight posterior distribution. MC dropout (Gal & Ghahramani, 2016) reinterprets the dropout (Srivastava et al., 2014) as an approximate VI, which retains the standard NN training procedure and modifies only the inference procedure for posterior MC approximation. In a similar spirit, some approaches (Mandt et al., 2017; Zhang et al., 2018; Maddox et al., 2019; Osawa et al., 2019) sequentially estimate the mean and covariance of the weight posterior distribution by using gradients computed at each step. As different from the BNNs, Deep kernel learning (Wilson et al., 2016a;b) places Gaussian processes (GPs) on top of the “deterministic” NNs, which combines NNs’ capability of handling complex high dimensional data and GPs’ capability of principled uncertainty representation and robust extrapolation.

Non-Bayesian approaches also help to resolve the limitations of softmax. Lakshminarayanan et al. (2017) propose an ensemble-based method to achieve better uncertainty representation and improved self-calibration. Both Guo et al. (2017) and Neumann et al. (2018) proposed temperature scaling-based methods for post-hoc modifications of softmax for improved calibration. To improve generalization by penalizing over-confidence, Pereyra et al. (2017) propose an auxiliary loss function that penalizes low predictive entropy, and Szegedy et al. (2016) and Xie et al. (2016) consider the types of noise included in ground-truth labels.

We also note that some recent studies use NNs to model the concentration parameter of the Dirichlet distribution but with a different purpose than BM. Sensoy et al. (2018) uses the loss function of explicitly minimizing prediction variances on training samples, which can help to produce high uncertainty prediction for out-of-distribution (OOD) or adversarial samples. Prior network (Malinin & Gales, 2018) investigates two types of auxiliary losses computed on in-distribution and OOD samples, respectively. Similar to prior network, Chen et al. (2018) considers an auxiliary loss computed on adversarially generated samples.

5. Experiment

In this section, we show versatility of BM through extensive empirical evaluations. We first verify its improvement of the generalization error in image classification tasks (section 5.1). We then verify whether BM inherits the advantages of the Bayesian method by placing the prior distribution *only* on the label categorical probability (section 5.2). We conclude this section by providing further applications that shows versatility of BM. To support reproducibility, we release our code at: <https://github.com/tjoo512/belief-matching-framework>. We performed all experiments on a single workstation with 8 GPUs (NVIDIA GeForce RTX 2080 Ti).

Throughout all experiments, we employ various large-scale models based on a residual connection (He et al., 2016a), which are the standard benchmark models in practice. For fair comparison and reducing burden of hyperparameter search, we fix experimental configurations to the reference implementation of corresponding architecture. However, we additionally use an initial learning rate warm-up and gradient clipping, which are extremely helpful for stable training of BM. Specifically, we use learning rates of $[0.1\epsilon, 0.2\epsilon, 0.4\epsilon, 0.6\epsilon, 0.8\epsilon]$ for first five epochs when the reference learning rate is ϵ and clip gradient when its norm exceeds 1.0. Without these methods, we had difficulty in training deep models, e.g., ResNet-50, due to gradient explosion at an initial stage of training.

We compare BM to following baseline methods: softmax, which is our primary object to improve; MC dropout with 100 MC samples, which is a simple and efficient BNN; deep ensemble with five NNs, which greatly improves the uncertainty representation ability. While there are other methods using NNs to model the Dirichlet distribution (Sensoy et al., 2018; Malinin & Gales, 2018; 2019), we note that these methods are not scalable to ResNet. Similarly, we observe that training a mixture of Dirichlet distributions (Wu et al., 2019b) with ResNet is subject to the gradient explosion, even with a 10x lower learning rate. Besides, BNNs with VI (or MCMC) are not directly comparable to our approach due to their huge modifications to existing baselines. For example, Heek & Kalchbrenner (2019) replace batch normalization and ReLU, use additional techniques (2x more filters, cyclic learning rate, snapshot ensemble), and require almost 10x more computations on ImageNet to converge.

5.1. Generalization Performance

We evaluate the generalization performance of BM on CIFAR (Krizhevsky, 2009) with the pre-activation ResNet (He et al., 2016b). CIFAR-10 and CIFAR-100 contain 50K training and 10K test images, and each 32x32x3-sized image belongs to one of 10 categories in CIFAR-10 and one of 100 categories in CIFAR-100. Table 1 lists the classification

Table 1. Test classification error rates on CIFAR. Here, we split a train set of 50K examples into a train set of 40K examples and a validation set of 10K example. Numbers indicate $\mu \pm \sigma$ computed across five trials, and boldface indicates the minimum mean error rate. Model and hyperparameter are selected based on validation error rates. We searched for the coefficients of BM over $\{0.01, 0.003, 0.001\}$ and MC dropout over $\{0.1, 0.2, 0.5\}$.

MODEL	METHOD	C-10	C-100
RES-18	SOFTMAX	6.13 \pm 0.13	26.44 \pm 0.33
	MC DROP (LAST)	6.13 \pm 0.08	26.15 \pm 0.10
	MC DROP (ALL)	6.50 \pm 0.14	27.32 \pm 0.45
	BM	5.93 \pm 0.07	24.19 \pm 0.34
RES-50	SOFTMAX	5.76 \pm 0.06	25.00 \pm 0.23
	MC DROP (LAST)	5.75 \pm 0.22	25.17 \pm 0.09
	MC DROP (ALL)	5.84 \pm 0.23	26.74 \pm 0.37
	BM	5.59 \pm 0.05	23.86 \pm 0.37

Table 2. Classification error rates on the ImageNet. Here, we use only $\lambda = 0.001$ for ResNext-50 and $\lambda = 0.0001$ for ResNext-101, and measure the validation error rates directly. We report the result obtained by single experiment due to computational constraint.

MODEL	METHOD	TOP1	TOP5
RESNEXT-50	SOFTMAX	22.23	6.36
	BM	22.03	6.29
RESNEXT-101	SOFTMAX	20.72	5.59
	BM	20.23	5.26

error rates of the softmax cross-entropy loss, BM, and MC-dropout. In all configurations, BM consistently achieves the best generalization performance on both datasets. On the other hand, last-layer MC dropout sometimes results in higher generalization errors than softmax and all-layer MC-dropout significantly increases error rates even though they consume 100x more computations for inference.

We next perform a large-scale experiment using ResNext-50 32x4d and ResNext-101 32x8d (Xie et al., 2017) on ImageNet (Russakovsky et al., 2015). ImageNet contains approximately 1.3M training samples and 50K validation samples, and each sample is resized to 224x224x3 and belongs to one of the 1K categories; that is, the ImageNet has more categories, a larger image size, and more training samples compared to CIFAR, which may enable a more precise evaluation of methods. Consistent with the results on CIFAR, BM improves test errors of softmax (Table 2). This result is appealing because improving the generalization error of deep NNs on large-scale datasets by adopting a Bayesian principle without computational overhead has rarely been reported in the literature.

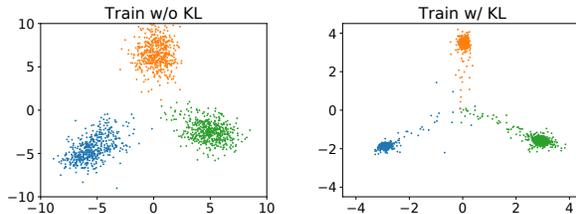


Figure 2. Penultimate layer’s activations of examples belonging to one of three classes (beaver, dolphin, and otter; indexed by 0,1,2 in CIFAR-100).

Regularization Effect of Prior In theory, BM has two regularization effects, which may explain the generalization performance improvements under BM: the prior distribution, which smooths the target posterior mean by adding pseudo counts, and computing the distribution matching loss by averaging of all possible categorical probabilities. In this regard, the ablation of the KL term in the ELBO helps to examine these two effects separately, which removes only the effect of the prior distribution.

We first examine its impact on the generalization performance by training a ResNet-50 on CIFAR without the KL term. The resulting test error rates were **5.68%** on CIFAR-10 and **24.69%** on CIFAR-100. These significant reductions in generalization performances indicates the powerful regularization effect of the prior distribution (cf. Table 1). The result that BM without the KL term still achieves lower test error rates compared to softmax demonstrates the regularization effect of considering all possible categorical probabilities by the Dirichlet distribution instead of choosing single categorical probability.

Considering the role of the prior distribution on smoothing the posterior mean, we conjecture that the impact of the prior distribution is similar to the effect of label smoothing. In Müller et al. (2019), it is shown that label smoothing makes learned representation reveal tight clusters of data points within the same classes and smaller deviations among the data points. Inspired by this result, we analyze the activations in the penultimate layer with the visualization method proposed in Müller et al. (2019). Figure 2 illustrates that the prior distribution significantly reduces the value ranges of the activations of data points, which implies the implicit function regularization effect considering that the L^p norm of $f^W \in L^p(\mathcal{X})$ can be approximated by $\|f^W\|_p \approx (\frac{1}{N} \sum_i |f^W(\mathbf{x}^{(i)})|^p)^{1/p}$. Besides, Figure 2 shows that the prior distribution makes activations belong to the same class to form much tighter clusters, which can be thought of as the implicit manifold regularization effect. To see this, assume that two images belonging to the same class have close distance in the data manifold. Then, the

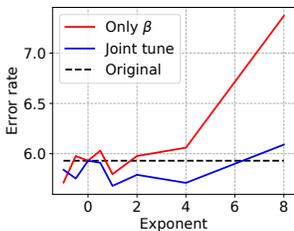


Figure 3. Impact of β on generalization performance. We exclude the ranges $\beta < \exp(-2)$ and $\beta > \exp(8)$ because the ranges result in gradient explosion under the strategy of changing only β

difference between logits of same class examples becomes a good proxy for the gradient of $f^{\mathbf{W}}$ along the data manifold since the gradient measures changes in the output space with respect to small changes in the input space.

Impact of β In section 3.4, we claimed that a value of β is implicitly related to the distribution of logit values, and its extreme value can be detrimental to the training stability. We verify this claim by training ResNet-18 on CIFAR-10 with different values of β . Specifically, we examine two strategies of changing β : modifying only β or jointly modifying the lambda proportional to $1/\beta$ to match local optima (cf. section 3.4). As a result, we obtain a robust generalization performance in both strategies when $\beta \in [\exp(-1), \exp(4)]$ (Figure 3). However, when β becomes extremely small ($\exp(-2)$ when changing only β and $\exp(-8)$ when jointly tuning λ and β), the gradient explosion occurs due to extreme slope of the digamma near 0. Conversely, when we increase only β to extremely large value, the error rate increases by a large margin (7.37) at $\beta = \exp(8)$, and eventually explodes at $\beta = \exp(16)$. This is because large beta increases the values of activations, so the gradient with respect to parameters explodes. Under the joint tuning strategy, such a high values region makes $\lambda \approx 0$, which removes the impact of the prior distribution.

5.2. Uncertainty Representation

One of the most attractive benefits of Bayesian methods is their ability to represent the uncertainty about their predictions. In a naive sense, uncertainty representation ability is the ability to “know what it doesn’t know.” For instance, models having a good uncertainty representation ability would increase some form of predictive uncertainty on misclassified examples compared to those on correctly classified examples. This ability is extremely useful in both real-world applications and downstream tasks in machine learning. For example, underconfident NNs can produce many false alarms, which makes humans ignore the predictions of NNs; conversely, overconfident NNs can exclude humans from the decision-making loop, which results in

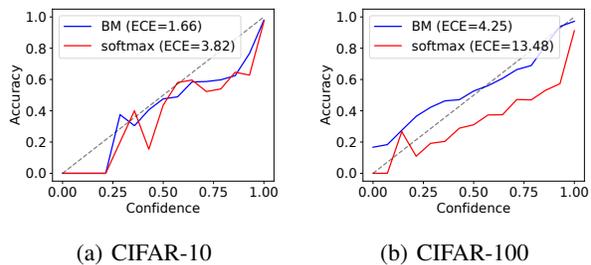


Figure 4. Reliability plots of ResNet-50 with BM and softmax. Here, ECE is computed with 15 groups.

catastrophic accidents. Also, better uncertainty representation enables balancing exploitation and exploration in reinforcement learning (Gal & Ghahramani, 2016) and detecting OOD samples (Malinin & Gales, 2018).

We evaluate the uncertainty representation ability on both in-distribution and OOD datasets. Specifically, we measure the calibration performance on in-distribution test samples, which examines a model’s ability to match its probabilistic output associated with an event to the actual long-term frequency of the event (Dawid, 1982). The notion of calibration in NNs is associated with how well its confidence matches the actual accuracy; e.g., we expect the average accuracy of a group of predictions having the confidence around 0.7 to be close to 70%. We also examine the predictive uncertainty for OOD samples. Since the examples belong to none of the classes seen during training, we expect neural networks to produce outputs of “I don’t know.”

In-Distribution Uncertainty We measure the calibration performance by the expected calibration error (ECE; Naeini et al., 2015), in which $\max_i \phi_i(f^{\mathbf{W}}(\mathbf{x}))$ is regarded as a prediction confidence for the input \mathbf{x} . ECE is calculated by grouping predictions based on the confidence score and then computing the absolute difference between the average accuracy and average confidence for each group; that is, the ECE of $f^{\mathbf{W}}$ on \mathcal{D} with M groups is as follows:

$$\text{ECE}_M(f^{\mathbf{W}}, \mathcal{D}) = \sum_{i=1}^M \frac{|\mathcal{G}_i|}{|\mathcal{D}|} |\text{acc}(\mathcal{G}_i) - \text{conf}(\mathcal{G}_i)| \quad (14)$$

where \mathcal{G}_i is a set of samples in the i -th group, defined as $\mathcal{G}_i = \{j : i/M < \max_k \phi_k(f^{\mathbf{W}}(\mathbf{x}^{(j)})) \leq (i+1)/M\}$, $\text{acc}(\mathcal{G}_i)$ is an average accuracy in the i -th group, and $\text{conf}(\mathcal{G}_i)$ is an average confidence in the i -th group.

We analyze the calibration property of ResNet-50 examined in section 5.1. As Figure 4 presents, BM’s predictive probability is well matched to its accuracy compared to softmax—that is, BM improves the calibration property of NNs. Specifically, BM improves ECE of softmax from 3.82 to 1.66 on CIFAR-10 and from 13.48 to 4.25 on CIFAR-100,

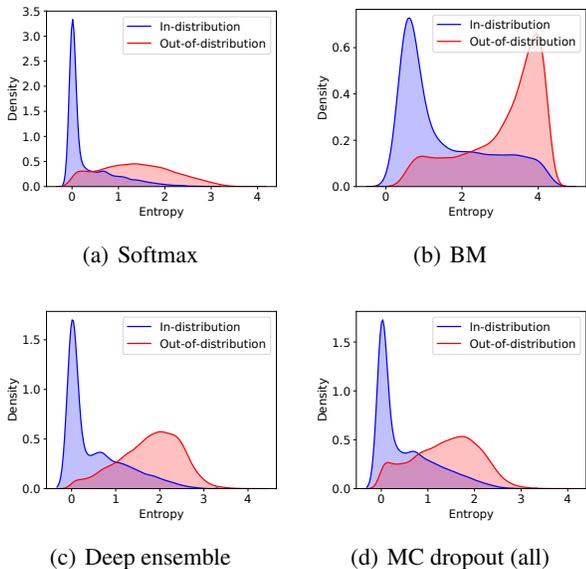


Figure 5. Uncertainty representation for in-distribution (CIFAR-100) and OOD (SVHN) of ResNet-50 under softmax and BM. We exclude the result of last-layer MC dropout because it shows no meaningful difference compared to the softmax.

respectively. These improvements are comparable to the deep ensemble, which achieves 1.04 on CIFAR-10 and 3.54 on CIFAR-100 with 5x more computations for both training and inference. In the case of all-layer MC dropout, ECE decreases to 1.50 on CIFAR-10 and 9.76 on CIFAR-100, however, these improvements require to compromise the generalization performance. On the other hand, the last-layer MC dropout, which often improves the generalization performance, does not show meaningful ECE improvements (3.78 on CIFAR-10 and 13.52 on CIFAR-100). We note that there are also post-hoc solutions improving calibration performance, e.g., temperature scaling (Guo et al., 2017). However, these methods often require an additional dataset for tuning the behavior of NNs, which may prevent the exploitation of entire samples to train NNs.

Out-of-Distribution Uncertainty We quantify uncertainty by predictive entropy, which measures the uncertainty of $f^{\mathbf{W}}(\mathbf{x})$ as follows: $H[\mathbb{E}_{q_{\mathbf{z}|\mathbf{x}}}[\mathbf{z}]] = H[\phi(f^{\mathbf{W}}(\mathbf{x}))] = -\sum_{k=1}^K \phi_k(f^{\mathbf{W}}(\mathbf{x})) \log \phi_k(f^{\mathbf{W}}(\mathbf{x}))$. This uncertainty measure gives intuitive interpretation such that the “I don’t know” the answer is close to the uniform distribution; conversely, “I confidently know” answer has one dominating categorical probability.

Figure 5 presents density plots of the predictive entropy, showing that BM provides notably better uncertainty estimation compared to other methods. Specifically, BM makes clear peaks of predictive entropy in high uncertainty re-

Table 3. Transfer learning performance (test error rates) from ResNet-50 pretrained on ImageNet to smaller datasets. μ and σ are obtained by five experiments, and boldface indicates the minimum mean error rate. We examine only $\lambda = 0.01$ for BM.

METHOD	C-10	FOOD-101	CARS
SOFTMAX	5.44 ± 0.10	28.49 ± 0.08	42.99 ± 0.14
BM	5.03 ± 0.04	26.41 ± 0.07	39.99 ± 0.20

gion for OOD samples (Figure 5(b)). In contrast, softmax produces relatively flat uncertainty for OOD samples (Figure 5(a)). Even though both MC dropout and deep ensemble successfully increase predictive uncertainty for OOD samples compared to softmax, they fail to make a clear peak on the high uncertainty region for such samples, unlike BM. We note that this remarkable result is obtained by being Bayesian only about the categorical probability.

Note that some in-distribution samples should be answered as “I don’t know” because the network does not achieve perfect test accuracy. As Figure 5 shows, BM contains more samples of high uncertainty for in-distribution samples compared to softmax that is almost certain in its predictions. This result consistently supports the previous result that BM resolves the overconfidence problem of softmax.

5.3. Transfer Learning

BM adopts the Bayesian principle *outside* the NNs, so it can be applied to models already trained on different tasks, unlike BNNs. In this regard, we examine the effectiveness of BM on the transfer learning scenario. Specifically, we downloaded the ImageNet-pretrained ResNet-50, and fine-tune weights of the last linear layer for 100 epochs by the Adam optimizer (Kingma & Ba, 2015) with learning rate of $3e-4$ on three different datasets (CIFAR-10, Food-101 (Bossard et al., 2014), and Cars (Krause et al., 2013)).

Table 3 compares test error rates of softmax and BM, in which BM consistently achieves better performances compared to softmax. Next, we examine the predictive uncertainty for OOD samples (Figure 6). Surprisingly, we observe that BM significantly improves the uncertainty representation ability of pretrained-models *by only fine-tuning the last layer weights*. These results present a possibility of adopting BM as a post-hoc solution to enhance the uncertainty representation ability of pretrained models without sacrificing their generalization performance. We believe that the interaction between BM and the pretrained models is significantly attractive, considering recent efforts of the deep learning community to construct general baseline models trained on extremely large-scale datasets and then transfer the baselines to multiple down-stream tasks (e.g., BERT (Devlin et al., 2018) and MoCo (He et al., 2019)).

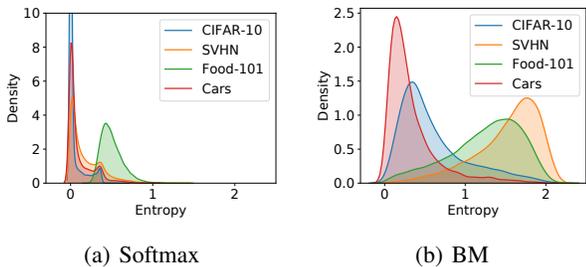


Figure 6. Uncertainty representation for in-distribution samples (CIFAR-10) and OOD samples (SVHN, Foods, and Cars) in transfer learning tasks. BM produces clear peaks in high uncertainty region on SVHN and Food-101. We note that BM confidently predicts examples in Cars because CIFAR-10 contains the object category of “automobile”. On the other hand, softmax produces confident predictions on all datasets compared to BM.

5.4. Semi-Supervised Learning

BM enables NNs to represent rich information in their predictions (cf. section 3.2). We exploit this characteristic to benefit consistency-based loss functions for semi-supervised learning. The idea of consistency-based losses employs information of unlabelled samples to determine where to promote robustness of predictions under stochastic perturbations (Belkin et al., 2006; Oliver et al., 2018). In this section, we investigate two baselines that consider stochastic perturbations on inputs (VAT; Miyato et al., 2018) and networks (II-model; Laine & Aila, 2017), respectively. Specifically, VAT generates adversarial direction \mathbf{r} , then measures KL-divergence between predictions at \mathbf{x} and $\mathbf{x} + \mathbf{r}$:

$$\mathcal{L}_{VAT}(\mathbf{x}) = KL(\phi(f^{\mathbf{W}}(\mathbf{x})) \parallel \phi(f^{\mathbf{W}}(\mathbf{x} + \mathbf{r}))) \quad (15)$$

where the adversarial direction is chosen by $\mathbf{r} = \arg \max_{\|\mathbf{r}\| \leq \epsilon} KL(\phi(f^{\mathbf{W}}(\mathbf{x})) \parallel \phi(f^{\mathbf{W}}(\mathbf{x} + \mathbf{r})))$, and II-model measures L^2 distance between predictions with and without enabling stochastic parts in NNs:

$$\mathcal{L}_{II}(\mathbf{x}) = \|\phi(\bar{f}^{\mathbf{W}}(\mathbf{x})) - \phi(f^{\mathbf{W}}(\mathbf{x}))\|_2^2 \quad (16)$$

where \bar{f} is a prediction without the stochastic parts.

We can see that both methods achieve the perturbation invariant predictions by minimizing the divergence between two categorical probabilities under some perturbations. In this regard, BM can provide a more delicate measure of the prediction consistency–divergence between Dirichlet distributions—that can capture richer probabilistic structures, e.g., (co)variances of categorical probabilities. This generalization the moment matching problem to the distribution matching problem can be achieved by replacing only the consistency measures in equation 15 with $KL(q_{\mathbf{z}|\mathbf{x}}^{\mathbf{W}} \parallel q_{\mathbf{z}|\mathbf{x}+\mathbf{r}}^{\mathbf{W}})$ and in equation 16 with $KL(\bar{q}_{\mathbf{z}|\mathbf{x}}^{\mathbf{W}} \parallel q_{\mathbf{z}|\mathbf{x}}^{\mathbf{W}})$.

Table 4. Classification error rates on CIFAR-10. μ and σ are obtained by five experiments, and boldface indicates the minimum mean error rate. We matched the configurations to those of Oliver et al. (2018) except for a consistency loss coefficient of 0.03 for VAT and 0.5 for II-model to match the scale between supervised and unsupervised losses. We use only $\lambda = 0.01$ for BM.

METHOD	II-MODEL	VAT
SOFTMAX	16.52 \pm 0.21	13.33 \pm 0.37
BM	16.01 \pm 0.36	12.40 \pm 0.23

We train wide ResNet 28-2 (Zagoruyko & Komodakis, 2016) via the consistency-based loss functions on CIFAR-10 with 4K/41K/5K number of labeled training/unlabeled training/validation samples. Our experimental results show that the distribution matching metric of BM is more effective than the moment matching metric under the softmax on reducing the error rates (Table 4). We think that the improvement in semi-supervised learning with more sophisticated consistency measures shows the potential usefulness of BM on other useful applications. This is because employing the prediction difference of neural networks is a prevalent method in many domains such as knowledge distillation (Hinton et al., 2015) and model interpretation (Zintgraf et al., 2017).

6. Conclusion

We adopted the Bayesian principle for constructing the target distribution by considering the categorical probability as a random variable rather than being given by the training label. The proposed method can be flexibly applied to the standard deep learning models by replacing only the softmax and the cross-entropy loss, which provides the consistent improvements in generalization performance, better uncertainty estimation, and well-calibrated behavior. We believe that BM shows promising advantages of being Bayesian about categorical probability.

We think that accommodating more expressive distributions in the belief matching framework is an interesting future direction. For example, parameterizing the logistic normal distribution (or mixture distribution) can make neural networks to capture strong semantic similarities among class labels, which would be helpful in large-class classification problems such as machine translation and classification on the ImageNet. Besides, considering the input dependent prior would result in interesting properties. For example, under the teacher-student framework, the teacher can dictate the prior for each input, thereby control the desired smoothness of the student’s prediction on the location. This property can benefit various domains such as imbalanced datasets and multi-domain learning.

Acknowledgements

We would like to thank Dong-Hyun Lee and anonymous reviewers for the discussions and suggestions.

References

- Amari, S.-I. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.
- Belkin, M., Niyogi, P., and Sindhwani, V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(Nov):2399–2434, 2006.
- Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural networks. In *International Conference on Machine Learning*, 2015.
- Bossard, L., Guillaumin, M., and Van Gool, L. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.
- Bridle, J. S. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, pp. 227–236. Springer, 1990.
- Chen, W., Shen, Y., Jin, H., and Wang, W. A variational Dirichlet framework for out-of-distribution detection. *arXiv preprint arXiv:1811.07308*, 2018.
- Dawid, A. P. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Gal, Y. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.
- Gal, Y. and Ghahramani, Z. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.
- Gong, W., Li, Y., and Hernández-Lobato, J. M. Meta-learning for stochastic gradient MCMC. In *International Conference on Learning Representations*, 2019.
- Graves, A. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, 2011.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.
- Hafner, D., Tran, D., Lillicrap, T., Irpan, A., and Davidson, J. Noise contrastive priors for functional uncertainty. *arXiv preprint arXiv:1807.09289*, 2018.
- He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, 2015.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016a.
- He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, 2016b.
- He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.
- Heek, J. and Kalchbrenner, N. Bayesian inference for large scale image classification. *arXiv preprint arXiv:1908.03491*, 2019.
- Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 2015.
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Machine Learning*, 2015.
- Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *ICCV Workshop on 3D Representation and Recognition*, pp. 554–561, 2013.
- Krizhevsky, A. Learning multiple layers of features from tiny images. *Technical report*, 2009.
- Laine, S. and Aila, T. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.
- Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

- LeCun, Y., Bottou, L., Orr, G. B., and Müller, K.-R. Efficient backprop. In *Neural Networks: Tricks of the Trade*, pp. 9–50. Springer, 1998.
- Louizos, C. and Welling, M. Multiplicative normalizing flows for variational Bayesian neural networks. In *International Conference on Machine Learning*, 2017.
- Ma, Y.-A., Chen, T., and Fox, E. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, 2015.
- MacKay, D. J. A practical Bayesian framework for back-propagation networks. *Neural Computation*, 4(3):448–472, 1992.
- MacKay, D. J. Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505, 1995.
- Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. A simple baseline for Bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, 2019.
- Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, 2018.
- Malinin, A. and Gales, M. Reverse KL-divergence training of prior networks: Improved uncertainty and adversarial robustness. In *Advances in Neural Information Processing Systems*, 2019.
- Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18(1):4873–4907, 2017.
- Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.
- Müller, R., Kornblith, S., and Hinton, G. When does label smoothing help? In *Advances in Neural Information Processing Systems*, 2019.
- Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015.
- Neumann, L., Zisserman, A., and Vedaldi, A. Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection. In *NIPS Workshop on Machine Learning for Intelligent Transportation Systems*, 2018.
- Oliver, A., Odena, A., Raffel, C. A., Cubuk, E. D., and Goodfellow, I. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in Neural Information Processing Systems*, 2018.
- Osawa, K., Swaroop, S., Jain, A., Eschenhagen, R., Turner, R. E., Yokota, R., and Khan, M. E. Practical deep learning with Bayesian principles. In *Advances in Neural Information Processing Systems*, 2019.
- Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., and Hinton, G. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.
- Raiko, T., Valpola, H., and LeCun, Y. Deep learning made easier by linear transformations in perceptrons. In *International Conference on Artificial Intelligence and Statistics*, 2012.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- Schraudolph, N. Accelerated gradient descent by factor-centering decomposition. *Technical report*, 1998.
- Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, 2018.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.
- Sun, S., Zhang, G., Shi, J., and Grosse, R. Functional variational Bayesian neural networks. In *International Conference on Learning Representations*, 2019.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, 2011.
- Wiesler, S., Richard, A., Schlüter, R., and Ney, H. Mean-normalized stochastic gradient for large-scale deep learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.
- Wilson, A. G., Hu, Z., Salakhutdinov, R. R., and Xing, E. P. Deep kernel learning. In *International Conference on Artificial Intelligence and Statistics*, 2016a.

- Wilson, A. G., Hu, Z., Salakhutdinov, R. R., and Xing, E. P. Stochastic variational deep kernel learning. In *Advances in Neural Information Processing Systems*, 2016b.
- Wu, A., Nowozin, S., Meeds, E., Turner, R. E., Hernández-Lobato, J. M., and Gaunt, A. L. Deterministic variational inference for robust Bayesian neural networks. In *International Conference on Learning Representations*, 2019a.
- Wu, Q., Li, H., Li, L., and Yu, Z. Quantifying intrinsic uncertainty in classification via deep dirichlet mixture networks. *arXiv preprint arXiv:1906.04450*, 2019b.
- Xie, L., Wang, J., Wei, Z., Wang, M., and Tian, Q. Disturblabel: Regularizing cnn on the loss layer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. Aggregated residual transformations for deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- Zagoruyko, S. and Komodakis, N. Wide residual networks. In *British Machine Vision Conference*, 2016.
- Zhang, G., Sun, S., Duvenaud, D., and Grosse, R. Noisy natural gradient as variational inference. In *International Conference of Machine Learning*, 2018.
- Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. Visualizing deep neural network decisions: Prediction difference analysis. In *International Conference on Learning Representations*, 2017.