EVI-BALD: BAYESIAN ACTIVE LEARNING BY DISAGREEMENT VIA EVIDENTIAL DEEP LEARNING

Minghao Li
Department of Computer Science and Engineering
University of North Texas
Denton, TX 76207, USA
MinghaoLi@my.unt.edu

Weishi Shi
Department of Computer Science and Engineering
University of North Texas
Denton, TX 76207, USA
Weishi.Shi@unt.edu

ABSTRACT

Bayesian Active Learning by Disagreement (BALD) is a fundamental acquisition function for active learning that measures the mutual information between model parameters and predictions. However, existing BALD computation methods rely on Monte Carlo sampling through ensembles or MC dropout, which require training and maintaining multiple independent models, leading to substantial computational overhead and time consumption. In this paper, we propose Evi-BALD, a hierarchical Bayesian approach that computes BALD efficiently using a single model instead of Monte Carlo sampling. Our method builds a second-order distribution over the predictive distribution, where the second-order parameters are directly predicted by the neural network, eliminating the need for multiple model instances. To address the computational challenge of evaluating intractable integrals in BALD calculation, we leverage Taylor expansion and demonstrate that BALD can be represented as the sum of quadratic terms and higher-order remainder terms from the entropy Taylor expansion. Through experiments on CIFAR-10, Evi-BALD achieves leading performance in active learning while significantly reducing computational time, demonstrating both efficiency and effectiveness.

1 INTRODUCTION

Supervised machine learning typically requires large amounts of labeled data, yet in many real-world applications labels are expensive, time-consuming, or difficult to obtain. Data-efficient learning methods are therefore of great practical interest. Among them, active learning reduces the need for labels by asking an oracle to label only a small, carefully chosen subset of the data. The key challenge is to decide which unlabeled examples are most worth labeling so that the model improves as much as possible with minimal annotation cost. Connecting this choice to the model's uncertainty over its predictions leads to uncertainty-based active learning. Active learning Cohn et al. (1996) aims to train accurate models with as few labeled examples as possible by iteratively selecting the most informative unlabeled samples for annotation. It is of broad importance in settings where labels are costly or scarce, such as medical imaging Hoi et al. (2006), scientific discovery Calinon et al. (2007), and low-resource domains. Meanwhile, quantifying uncertainty in machine learning is essential. Models are not only wrong or right but often uncertain, and the model uncertainty is an essential sign for model understanding, improvement and robustness. It indicates where the model would benefit more from additional data Yu et al. (2024) or is vulnerable to the out-of-distribution (OOD) problems. In particular, active learning can be viewed as a targeted mechanism for coping with distributional drift, because it continually updates the model by querying labels for newly encountered, potentially shifted data.

Bayesian Active Learning by Disagreement (BALD) Houlsby et al. (2011); Gal et al. (2017); Kirsch et al. (2019); Beluch et al. (2018) is different from the other AL methods. It measures the impact of observing a new data sample on the models. Currently, the BALD AL method relies on sampling methodologies like Ensemble Beluch et al. (2018); Lakshminarayanan et al. (2017) by using different initializations or adversarial training. MC Dropout Gal & Ghahramani (2016), where the uncertainty in the weights induces prediction uncertainty by marginalizing over the approximate posterior using Monte Carlo integration, also creates a group of models with random Dropout and

then measures the uncertainty information. These Monte Carlo sampling based methods highly rely on the behavior of the Ensemble form.

Recently people have found the downside of the ensemble-based method. Abe et al. Abe et al. (2022) show that ensemble diversity, by any metric, does not meaningfully contribute to an ensemble's uncertainty quantification on OOD data, but is instead highly correlated with the relative improvement of a single larger model. With limited computation power, the ensemble also breaks the original algorithm structure, since it separates the training resources into multiple independent parts. Then a natural question is how to measure the BALD score without the ensemble or MC Dropout. A hierarchical Bayesian based method can be a solution.

A central element of Bayesian inference is the assignment of prior distributions in hierarchical models, which enables rigorous quantification and estimation of uncertainty (Gelman, 2006). The Evidential Deep Learning (EDL) framework (Sensoy et al., 2018) addresses predictive uncertainty in multi-class classification by parameterizing a Dirichlet prior over the categorical output distribution. A KL divergence EDL form builds a second-order distribution directly on one-hot encoded ground truth Malinin & Gales (2018; 2019), aiming to encourage a diffused prior for out-of-distribution data, and a more concentrated Dirichlet distribution for in-distribution data. An ELBO form of a second-order predictor maximizes a lower bound on the logarithm of the marginal likelihood of the observations Charpentier et al. (2021; 2020); Chen et al. (2018).

EDL related methods give support for a deeper analysis of a specific model uncertainty instead of building a bag of models. It can be seen as a hierarchical Bayesian model with a second-order distribution on the model parameters. In this paper, we propose the Evi-BALD to approximate the BALD score, which is efficient and flexible. Specifically, we adopt Type-II maximum likelihood to build a second-order distribution directly predicted by a single neural network, eliminating the need for Monte Carlo sampling from multiple models. To address the computational challenge of evaluating the intractable expected entropy term, we derive a second-order Taylor expansion approximation that reveals BALD as the sum of quadratic variance terms and higher-order remainder terms. This formulation enables efficient BALD computation while maintaining theoretical rigor and practical applicability across different model architectures.

2 BACKGROUND

2.1 BAYESIAN ACTIVE LEARNING

In active learning for Bayesian models, a key goal is to select data points from a pool that are expected to provide the most valuable information for refining the model's understanding of the underlying data distribution. Specifically, we aim to choose points that maximize the information gain about the model parameters $\boldsymbol{\theta}$, which are the weights or hyperparameters of the model. This is formalized by maximizing the mutual information $I(\mathbf{t}; \boldsymbol{\theta} \mid \mathbf{x}, \mathcal{D})$ between the prediction targets $\mathbf{t}$ for a new input $\mathbf{x}$ and the posterior distribution over the model parameters $p(\boldsymbol{\theta} \mid \mathcal{D})$, given the training data $\mathcal{D}$.

BALD is a widely recognized and highly effective acquisition function in Bayesian active learning, as it quantifies epistemic uncertainty by measuring mutual information between predictions and model parameters. The mutual information from the BALD Houlsby et al. (2011) can be expressed as:

$$I(\mathbf{t}; \boldsymbol{\theta} \mid \mathbf{x}, \mathcal{D}) = H(\mathbf{t} \mid \mathbf{x}, \mathcal{D}) - \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D})} \left[H(\mathbf{t} \mid \mathbf{x}, \boldsymbol{\theta}) \right],$$

High mutual information indicates that observing $\mathbf{t}$ would update our beliefs about $\boldsymbol{\theta}$, resolving ambiguity in the parameters by distinguishing between competing models that predict conflicting outcomes with high confidence.

2.2 BALD UNDER TYPE I MAXIMUM LIKELIHOOD

In maximum likelihood estimation (MLE) for classification, the goal is to optimize $\boldsymbol{\theta}$, which we identify with the model parameters $\boldsymbol{\theta}$, to maximize $p(\mathcal{D} \mid \boldsymbol{\theta}) = \prod_i p(\mathbf{y}_i \mid \mathbf{x}_i, \boldsymbol{\theta})$. For simplicity, we will not note the i for the i-th sample. We model a continuous hyperparameter vector $\boldsymbol{\mu}$ as the output of a stochastic model $f(\boldsymbol{\mu} \mid \boldsymbol{\theta}, \mathbf{x})$, yielding a surrogate likelihood $p(\mathcal{D} \mid \boldsymbol{\mu})$. Here, $\boldsymbol{\theta}$ is "ab-

sorbed" into $\boldsymbol{\mu}$, meaning $\boldsymbol{\mu}$ parameterizes the predictive distribution $p(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\mu})$. This formulation enables gradient-based optimization and provides a built-in first-order confidence measure via $\boldsymbol{\mu}$, often referred to as a level-1 predictor Bengs et al. (2022).

In a Bayesian active learning context, we can extend this by placing a posterior over $\boldsymbol{\mu}$ induced by the posterior over $\boldsymbol{\theta}$, where Gal et al. (2017) swaps the posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$ in $p(\boldsymbol{\mu} \mid \mathbf{x}, \mathcal{D}) = \int p(\boldsymbol{\mu} \mid \mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D})\, d\boldsymbol{\theta}$ with the approximated posterior by Monte Carlo sampling. Since $\boldsymbol{\mu} = f(\mathbf{x}, \boldsymbol{\theta})$ is a deterministic function of $\boldsymbol{\theta}$ for a fixed $\mathbf{x}$, the entropies and expectations can be rewritten in terms of $\boldsymbol{\mu}$. The mutual information can then be represented by:

$$I(\mathbf{t}; \boldsymbol{\mu} \mid \mathbf{x}, \mathcal{D}) \; = H(\mathbf{t} \mid \mathbf{x}, \mathcal{D}) - \mathbb{E}_{p(\boldsymbol{\mu} \mid \mathbf{x}, \mathcal{D})} \left[H(\mathbf{t} \mid \boldsymbol{\mu}, \mathbf{x}) \right], \tag{1}$$

The second term is the expectation of the conditional entropy with respect to the posterior over parameters. Because of the conditional independence $\mathbf{t} \perp \mathcal{D} \mid \boldsymbol{\mu}, \mathbf{x}$, we have

$$H(\mathbf{t} \mid \boldsymbol{\mu}, \mathbf{x}, \mathcal{D}) = H(\mathbf{t} \mid \boldsymbol{\mu}, \mathbf{x}).$$

The predictive distribution depends on $\boldsymbol{\mu}$ and $\mathbf{x}$ but not directly on the training data once $\boldsymbol{\mu}$ is known.

One of the challenges for BALD under the Type I maximum likelihood is approximating the expectation of the model hyperparameters. For example, the $p(\mathbf{t} \mid \mathbf{x}, \mathcal{D})$ integrates the model stochastic parameters $\boldsymbol{\theta}$ and hyperparameter $\boldsymbol{\mu}$:

$$\begin{aligned} p(\mathbf{t} \mid \mathbf{x}, \mathcal{D}) &= \mathbb{E}_{\boldsymbol{\mu}} \left[\mathbb{E}_{\boldsymbol{\theta}} \left[p(\mathbf{t} \mid \boldsymbol{\mu}, \mathbf{x}, \boldsymbol{\theta}, \mathcal{D}) \right] \right] \\ &= \int \int p(\mathbf{t} \mid \boldsymbol{\mu}, \boldsymbol{\theta}, \mathbf{x}, \mathcal{D}) d\boldsymbol{\theta} d\boldsymbol{\mu} \\ &= \int \int p(\mathbf{t} \mid \boldsymbol{\mu}) p(\boldsymbol{\mu} \mid \mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D})\, d\boldsymbol{\theta} d\boldsymbol{\mu} \end{aligned}$$

For the expectation with respect to $\boldsymbol{\mu}$, MC dropout and Ensemble give different solutions by the Monte Carlo integration. But The standard approach to computing BALD relies on Monte Carlo sampling from the posterior, which typically requires training and maintaining an ensemble of multiple independent models, which is more resource-consuming.

2.3 Maximizing Type II Likelihood

Type-II maximum likelihood estimation (MLE), also known as the evidence approximation MacKay (1992) or empirical Bayes, provides a widely used Bayesian approximation strategy. Given a predictive target $\mathbf{t}$, first-order hyperparameters $\boldsymbol{\mu}$, and second-order hyperparameters $\mathbf{m}$, the method optimizes $\mathbf{m}$ by maximizing the marginal likelihood $p(\mathcal{D} \mid \mathbf{m})$ under the following assumption: The posterior $p(\mathbf{m} \mid \mathcal{D})$ is sharply concentrated around its mode $\mathbf{m}^*$, while the prior $p(\mathbf{m})$ is non-informative.

This sharply peaked posterior justifies approximating it with a Dirac delta at the mode: $p(\mathbf{m} \mid \mathcal{D}) \approx \delta(\mathbf{m} - \mathbf{m}^*)$. Substituting this approximation greatly simplifies the predictive distribution:

$$p(\mathbf{t} \mid \mathbf{x}, \mathcal{D}) = \iint p(\mathbf{t} \mid \boldsymbol{\mu})\, p(\boldsymbol{\mu} \mid \mathbf{m})\, p(\mathbf{m} \mid \mathbf{x}, \mathcal{D})\, \mathrm{d}\boldsymbol{\mu}\, \mathrm{d}\mathbf{m} \approx \int p(\mathbf{t} \mid \boldsymbol{\mu})\, p(\boldsymbol{\mu} \mid \mathbf{m}^*)\, \mathrm{d}\boldsymbol{\mu}.$$

A non-informative prior $p(\mathbf{m})$ ensures that the posterior is dominated by the likelihood term $p(\mathcal{D} \mid \mathbf{m})$, so that the mode $\mathbf{m}^*$ can be found by maximizing the Type-II marginal likelihood.

The Type-II likelihood is obtained by marginalizing out the lower-level parameters $\boldsymbol{\mu}$:

$$p(\mathcal{D} \mid \mathbf{m}) = \int p(\mathcal{D} \mid \boldsymbol{\mu})\, p(\boldsymbol{\mu} \mid \mathbf{m})\, \mathrm{d}\boldsymbol{\mu} = \mathbb{E}_{\boldsymbol{\mu} \sim p(\boldsymbol{\mu} \mid \mathbf{m})} \big[p(\mathcal{D} \mid \boldsymbol{\mu}) \big].$$

This marginal likelihood $p(\mathcal{D} \mid \mathbf{m})$ quantifies how well the hyperparameter setting $\mathbf{m}$ explains the observed dataset $\mathcal{D}$. Type-II maximum likelihood estimation then proceeds by solving

$$\hat{\mathbf{m}} = \arg\max_{\mathbf{m}} \; p(\mathcal{D} \mid \mathbf{m}).$$

By implicitly integrating out the model parameters $\boldsymbol{\mu}$, this procedure automatically selects hyperparameters that improve overall model fit and generalization.

3 BALD UNDER TYPE II MAXIMUM LIKELIHOOD

Ensemble and Dropout for BALD need to train multiple models to evaluate the second-order distribution by sampling. They highly rely on the structure of the group and also need more time for training a group of models.

To flexibly calculate the BALD on different model structures and save the training time, we propose Evi-BALD by Bayesian inference to build a second-order distribution on a single model. It supports fewer calculations and a mathematically compact way for solving the expectation with respect to $\boldsymbol{\mu}$ by using the Type II Maximum Likelihood.

Firstly, we can set a second-order distribution with hyperparameter $\mathbf{m}$. With the assumption of $p(\mathbf{m} \mid \mathbf{x}, \mathcal{D}) \approx \delta(\mathbf{m} - \mathbf{m}^*)$ and the Empirical Bayes Robbins (1992), the left term in equation 1 is defined as

$$\begin{aligned} H\left[p(\mathbf{t} \mid \mathbf{x}, \mathcal{D})\right] &= H\left[\int\int\int p(\mathbf{t} \mid \boldsymbol{\mu}, \mathbf{x}, \mathbf{m}, \boldsymbol{\theta}, \mathcal{D}) d\boldsymbol{\theta} d\boldsymbol{\mu} dm\right] \\ &= H\left[\int\int\int p(\mathbf{t} \mid \boldsymbol{\mu}) p(\boldsymbol{\mu} \mid \mathbf{m}) p(\mathbf{m} \mid \mathbf{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D})\, d\boldsymbol{\theta} d\boldsymbol{\mu} d\mathbf{m}\right] \\ &= H\left[\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m})}\left[\mathbb{E}_{\mathbf{m}\sim p(\mathbf{m}|\mathbf{x},\boldsymbol{\theta})}\left[\mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D})}\left[p(\mathbf{t} \mid \boldsymbol{\mu}, \mathbf{m}, \boldsymbol{\theta}, \mathbf{x}, \mathcal{D})\right]\right]\right]\right] \\ &= H\left[\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m})}\left[\mathbb{E}_{\mathbf{m}\sim p(\mathbf{m}|\mathbf{x},\mathcal{D})}\left[p(\mathbf{t} \mid \boldsymbol{\mu}, \mathbf{m}, \mathbf{x}, \mathcal{D})\right]\right]\right] \\ &\approx H\left[\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[p(\mathbf{t} \mid \boldsymbol{\mu}, \mathbf{m}^*, \mathbf{x}, \mathcal{D})\right]\right]. \end{aligned}$$

And the expectation over the entropy, the right term in equation 1, is defined as

$$\begin{aligned} \mathbb{E}_{p(\boldsymbol{\mu}|\mathbf{x},\mathcal{D})}\left[H(\mathbf{t} \mid \boldsymbol{\mu})\right] &= \mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m})}\left[\mathbb{E}_{\mathbf{m}\sim p(\mathbf{m}|\mathbf{x},\boldsymbol{\theta})}\left[\mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\mathcal{D})}\left[H(\mathbf{t} \mid \boldsymbol{\mu}, \mathbf{m}, \boldsymbol{\theta}, \mathbf{x}, \mathcal{D})\right]\right]\right] \\ &= \mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m})}\left[\mathbb{E}_{\mathbf{m}\sim p(\mathbf{m}|\mathbf{x},\mathcal{D})}\left[H(\mathbf{t} \mid \boldsymbol{\mu}, \mathbf{m}, \mathbf{x}, \mathcal{D})\right]\right] \\ &\approx \mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[H(\mathbf{t} \mid \boldsymbol{\mu}, \mathbf{m}^*, \mathbf{x}, \mathcal{D})\right]. \end{aligned}$$

For simplicity we note $H\left[\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[p(\mathbf{t} \mid \boldsymbol{\mu}, \mathbf{m}^*, \mathbf{x}, \mathcal{D})\right]\right]$ as $H\left[\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[p(\mathbf{t} \mid \boldsymbol{\mu})\right]\right]$ and $\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[H(\mathbf{t} \mid \boldsymbol{\mu}, \mathbf{m}^*, \mathbf{x}, \mathcal{D})\right]$ as $\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[H(\mathbf{t} \mid \boldsymbol{\mu})\right]$. In a general form of Equation equation 1 under Type II Maximum Likelihood, the mutual information is defined as

$$\begin{aligned} I(\mathbf{t}; \boldsymbol{\mu} \mid \mathbf{x}, \mathcal{D}) &= H(\mathbf{t} \mid \mathbf{x}, \mathcal{D}) - \mathbb{E}_{p(\boldsymbol{\mu}|\mathbf{x},\mathcal{D})}\left[H(\mathbf{t} \mid \mathbf{x}, \boldsymbol{\mu})\right] \\ &\approx H\left[\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[p(\mathbf{t} \mid \boldsymbol{\mu})\right]\right] - \mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[H(\mathbf{t} \mid \boldsymbol{\mu})\right]. \end{aligned} \tag{2}$$

Proposition 3.1. *Besides Monte Carlo sampling methods that rely on multiple models, BALD can be achieved using a single hierarchical Bayesian model. By applying Type-II maximum likelihood to infer the approximate posterior $p(\boldsymbol{\mu} \mid \mathbf{m}^*, \mathbf{x}, \mathcal{D})$, the BALD acquisition function is computed via tractable expectations over this single-model posterior.*

The expectation of the entropy $\mathbb{E}_{\boldsymbol{\mu}}[H(\mathbf{t} \mid \boldsymbol{\mu})]$ is generally intractable. A natural approach is to approximate it via Taylor expansion of the entropy functional around the mean predictive distribution $\mathbb{P}_0$, where $\mathbb{P}_0 = \mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m}^*,\mathbf{x},\mathcal{D})}\big[p(\mathbf{t} \mid \boldsymbol{\mu})\big]$ is the marginal distribution. The differential Shannon entropy is a strictly concave and infinitely differentiable function. These properties guarantee that the Taylor series exists and that higher-order remainder terms can be controlled in principle. We therefore write the second-order Taylor expansion of the entropy around $\mathbb{P}_0$:

$$H(\mathbb{P}) = H(\mathbb{P}_0) + \nabla H(\mathbb{P}_0)^\top(\mathbb{P} - \mathbb{P}_0) + \frac{1}{2}(\mathbb{P} - \mathbb{P}_0)^\top \nabla^2 H(\mathbb{P}_0)(\mathbb{P} - \mathbb{P}_0) + R_3(\mathbb{P}, \mathbb{P}_0),$$

where $\mathbb{P} = p(\mathbf{t} \mid \boldsymbol{\mu}, \mathbf{x}, \mathcal{D})$ is the predictive distribution induced by a draw of $\boldsymbol{\mu}$. $\nabla H(\mathbb{P}_0)$ is the gradient of the entropy functional with respect to the probability vector. $\nabla^2 H(\mathbb{P}_0)$ is the Hessian matrix of H evaluated at $\mathbb{P}_0$. $R_3(\mathbb{P}, \mathbb{P}_0)$ denotes the remainder term of the third and higher orders. Because the entropy is infinitely differentiable on the interior of the simplex and the predictive distributions are typically well away from the boundary in well-calibrated models, the remainder can often be made small or neglected in the vicinity of $\mathbb{P}_0$. We will discuss the potential limitation of the neglect of the remainder in the limitation section.

After taking the expectation on $\boldsymbol{\mu}$, it comes to:

$$\mathbb{E}_{\boldsymbol{\mu}}\big[H(\mathbb{P})\big] = H(\mathbb{P}_0) + \frac{1}{2}\operatorname{tr}\left[\nabla^2 H(\mathbb{P}_0)\cdot \operatorname{Cov}_{\boldsymbol{\mu}}(\mathbb{P})\right] + \mathbb{E}_{\boldsymbol{\mu}}[R_3(\mathbb{P},\mathbb{P}_0)], \tag{3}$$

where $\operatorname{Cov}_{\boldsymbol{\mu}}(\mathbb{P})$ is the $K\times K$ covariance matrix of the predictive probabilities over the posterior on $\boldsymbol{\mu}$, $\nabla^2 H(\mathbb{P}_0)$ is the Hessian of the categorical entropy at $\mathbb{P}_0$, which is diagonal with entries $-\frac{1}{(\mathbb{P}_0)_k}$. So the quadratic term simplifies enormously:

$$\frac{1}{2}\operatorname{tr}\left[\nabla^2 H(\mathbb{P}_0)\cdot \operatorname{Cov}_{\boldsymbol{\mu}}(\mathbb{P})\right] = -\frac{1}{2}\sum_{k=1}^{K}\frac{\operatorname{Var}_{\boldsymbol{\mu}}(\mathbb{P}_k)}{(\mathbb{P}_0)_k}$$

The remainder term $\mathbb{E}_{\boldsymbol{\mu}}[R_3(\mathbb{P},\mathbb{P}_0)]$ is mathematically complex because it involves higher-order integration over the distribution of $\boldsymbol{\mu}$. In our method, we only use the second-order Taylor expansion and neglect this remainder for tractability.

Now, when we look back at the mutual information in equation 2, the entropy of the expectation is canceled with the zeroth-order term. And the mutual information is equal to the higher-order terms of the Taylor expansion around the mean $\mathbb{P}_0$.

$$\begin{aligned} I(\mathbf{t};\boldsymbol{\mu}\mid \mathbf{x},\mathcal{D}) &\approx H\left[\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}\mid \mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[p(\mathbf{t}\mid\boldsymbol{\mu})\right]\right] - \mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}\mid \mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[H(\mathbf{t}\mid\boldsymbol{\mu})\right]. \\ &= \frac{1}{2}\sum_{k=1}^{K}\frac{\operatorname{Var}_{\boldsymbol{\mu}}(\mathbb{P}_k)}{(\mathbb{P}_0)_k} - \mathbb{E}_{\boldsymbol{\mu}}[R_3(\mathbb{P},\mathbb{P}_0)] \end{aligned} \tag{4}$$

Proposition 3.2. *Under Type-II MLE, the BALD mutual information $I(\mathbf{t};\boldsymbol{\mu}\mid\mathbf{x},\mathcal{D})$ approximates the negative expected conditional entropy after canceling the predictive entropy term. It thus equals the sum of quadratic and higher-order remainder terms in the Taylor expansion of the entropy around the mean predictive distribution $\mathbb{P}_0$, excluding the zeroth-order contribution.*

4 Evi-BALD for Classification

In the previous section, we demonstrated that the BALD acquisition function can be approximated using a second-order Taylor expansion of the entropy functional, leveraging Type-II MLE in place of traditional Monte Carlo sampling over the posterior. While this approach reduces computational cost, the resulting marginal predictive distribution under Type-II MLE renders exact integration for the expected entropy intractable. To address this challenge, we introduced a tractable second-order approximation of the entropy around the mean predictive distribution $\mathbb{P}_0$, yielding a general form expression for the Evi-BALD acquisition function.

Building on this foundation, the present section specializes the Evi-BALD framework to the practically important cases of binary and multi-class classification. We derive explicit forms of the acquisition function for each setting.

4.1 Binary Case

We will choose the Beta distribution, the conjugate prior of the Bernoulli distribution, as the second-order distribution for the Type II MLE Li et al.. Then the Entropy of the expectation of the Bernoulli distribution is derived by:

$$H\left[\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}\mid \mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[p(\mathbf{t}\mid\boldsymbol{\mu})\right]\right] = -\frac{a^*}{a^*+b^*}\log\frac{a^*}{a^*+b^*} - \frac{b^*}{a^*+b^*}\log\frac{b^*}{a^*+b^*}$$

where we substitute a^* and b^* for $\mathbf{m}^*$. Since $t\in\{0,1\}$ is fixed, the expectation of the Bernoulli likelihood factorizes into powers of the posterior mean of μ and $1-\mu$. And for cases $t=1$ and $t=0$, the expressions are identical, due to symmetry.

The expectation of the Entropy here is intractable, so we approximate it by the Taylor expansion of the entropy around the mean $\mathbb{E}[\mu]$ in Appendix A. The approximation will read:

$$\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}\mid \mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[H(\mathbf{t}\mid\boldsymbol{\mu})\right] \approx -\frac{a^*}{a^*+b^*}\log\frac{a^*}{a^*+b^*} - \frac{b^*}{a^*+b^*}\log\frac{b^*}{a^*+b^*} - \frac{1}{2(a^*+b^*+1)}.$$

Due to the complexity of the remainder terms of third order and higher, we will not specifically explore them here. In the end, we can get:

$$I(\mathbf{t}; \boldsymbol{\mu} \mid \mathbf{x}, \mathcal{D}) \approx \frac{1}{2(a^* + b^* + 1)}, \tag{5}$$

which is equal to the negative second-order of the Taylor expansion term $-\frac{1}{2}h''(\mu_0)\text{Var}(\mu)$ and the first-order term is canceled during the expectation.

4.2 Categorical Case

In the Categorical case, we will choose the Dirichlet distribution, the conjugate prior of the categorical distribution, as the second-order distribution for the Type II MLE Sensoy et al. (2018). We will replace the second-order parameters $\mathbf{m}$ with $\boldsymbol{\alpha}$ in equation 2. For simplicity we directly utilize the conclusion from equation 4 and it reads:

$$I(\mathbf{t}; \boldsymbol{\mu} \mid \mathbf{x}, \mathcal{D}) \approx \frac{K-1}{2(\alpha_0 + 1)} - \mathbb{E}_{\boldsymbol{\mu}}[R_3(\mathbb{P}, \mathbb{P}_0)],$$

where $\alpha_0 = \sum_{j=1}^{K} \alpha_j$. Additionally, we also derive it step by step. More details can be found in Appendix B. The entropy of the expectation of the Categorical distribution is derived as follows:

$$H\left[\mathbb{E}_{\boldsymbol{\mu} \sim p(\boldsymbol{\mu} \mid \mathbf{m}^*, \mathbf{x}, \mathcal{D})}\left[p(\mathbf{t} \mid \boldsymbol{\mu})\right]\right] = -\sum_{k=1}^{K} \frac{\alpha_k^*}{\alpha_0^*} \log \frac{\alpha_k^*}{\alpha_0^*},$$

where we substitute $\boldsymbol{\alpha}^*$ for $\mathbf{m}^*$. Since $\mathbf{t} \in \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$ is a fixed one-hot vector (with $t_{k^*} = 1$ for some class k^* and 0 elsewhere), the expectation of the Categorical likelihood factorizes into powers of the posterior means of the μ_k. Again, using the Taylor expansion for the expectation of the entropy, we obtain:

$$\mathbb{E}_{\boldsymbol{\mu} \sim p(\boldsymbol{\mu} \mid \mathbf{m}^*, \mathbf{x}, \mathcal{D})}\left[H(\mathbf{t} \mid \boldsymbol{\mu})\right] \approx -\sum_{k=1}^{K} \frac{\alpha_k^*}{\alpha_0^*} \log \frac{\alpha_k^*}{\alpha_0^*} - \frac{K-1}{2(\alpha_0^* + 1)}.$$

The final result is defined as

$$\begin{aligned} I(\mathbf{t}; \boldsymbol{\mu} \mid \mathbf{x}, \mathcal{D}) &= H(\mathbf{t} \mid \mathbf{x}, \mathcal{D}) - \mathbb{E}_{p(\boldsymbol{\mu} \mid \mathbf{x}, \mathcal{D})}\left[H(\mathbf{t} \mid \boldsymbol{\mu}, \mathbf{x})\right], \\ &\approx H\left[\mathbb{E}_{\boldsymbol{\mu} \sim p(\boldsymbol{\mu} \mid \mathbf{m}^*, \mathbf{x}, \mathcal{D})}\left[p(\mathbf{t} \mid \boldsymbol{\mu})\right]\right] - \mathbb{E}_{p(\boldsymbol{\mu} \mid \mathbf{m}^*, \mathbf{x}, \mathcal{D})}\left[H(\mathbf{t} \mid \boldsymbol{\mu})\right] \\ &\approx \frac{K-1}{2(\alpha_0 + 1)} \end{aligned} \tag{6}$$

where the α_0 is the sum of the Dirichlet hyperparameters.

5 Experiment

BALD is one of the most widely adopted and effective acquisition functions in modern Bayesian active learning. To thoroughly evaluate our proposed Evi-BALD method, we begin with active learning experiments. We first conduct experiments using multilayer perceptron (MLP) models on the CIFAR-10 dataset, a standard benchmark that allows us to assess performance in a controlled yet realistic setting.

Furthermore, a key contribution of our work is the introduction of a hierarchical Bayesian approach that computes the BALD score without relying on Monte Carlo sampling techniques, such as deep ensembles or MC dropout. This design eliminates the need to train and maintain multiple independent models, thereby substantially reducing computational cost and training time. For instance, while conventional Monte Carlo-based methods require training an ensemble of models to approximate posterior predictive uncertainty, our single-model hierarchical formulation relies solely on Type-II maximum likelihood to infer an approximate posterior distribution. In the experiments that follow, we explicitly quantify these time savings and demonstrate the practical efficiency gains of Evi-BALD compared to sampling-based baselines.

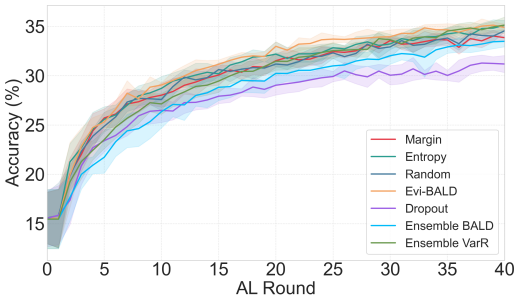

Figure 1: The proposed EVI-BALD approach demonstrated leading performance across six independent active learning experiments conducted on CIFAR-10.

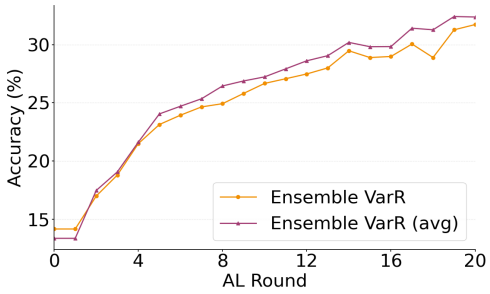

Figure 2: Ablation study on ensemble-based methods. We compare the performance using the first model versus the average over the ensemble. The ensemble average consistently outperforms the single model.

5.1 Active Learning

We run AL on CIFAR-10. The model is an evidential MLP and 10-class Dirichlet outputs. We use 10 initially labeled samples, balanced across classes, and 20 AL rounds. In each round we query 32 samples to be labeled, so the labeled set grows from 10 to 1290. In each round we train for 50 epochs with a batch size 32. For the Ensemble strategies, we set the number of independent models to 5. Abe et al. (2022) shows that the predictive power of an ensemble is highly correlated with the improvement of a single larger model. For fairness, at test time, we only use the prediction over the first model of the ensemble. We still use the whole ensemble to compute the query score, but only use the first model in ensemble for the evaluation instead of five. To better express the difference of it, we also conduct a quick ablation study experiment of this changing during the evaluation time. As shown in the Figure 2, we can see that the performance of the average of the ensemble is better than that of using only the first model. This is because the average of the ensemble has a similar performance as a bigger scale model Abe et al. (2022).

Figure 1 presents the performance of our proposed Evi-BALD method throughout the active learning process. The curve shows the mean accuracy, with the light-colored shaded area indicating the variance across six runs. The proposed Evi-BALD method achieved leading performance among multiple independent AL experiments. The comparison strategies are Entropy Lewis (1995), Margin Roth & Small (2006), Random, Dropout Gal et al. (2017), Ensemble BALD Beluch et al. (2018) and Ensemble VarR Beluch et al. (2018).

5.2 Time Efficiency

As illustrated in Figure 3, which displays the average per-round computational time across the active learning experiments, our proposed Evi-BALD method, shown in blue, achieves a substantial reduction compared to standard BALD implementations. This efficiency gain stems from its use of a hierarchical Bayesian model that approximates the posterior predictive distribution directly, eliminating the need for costly Monte Carlo sampling procedures in deep ensembles for BALD.

Table 1: Computational time (seconds) of active learning methods.

Method	Total	Train	Eval	Query	Per Round
Ensemble BALD	760.92	419.05	45.52	293.74	38.05
Ensemble VarR	1016.08	427.00	47.70	537.69	50.80
Dropout	382.73	74.93	39.72	265.89	19.14
Evi-BALD (Ours)	348.83	82.79	44.23	221.54	17.44

The timing summary table 1 presents time experiments compared with ensemble-based methods. Total is the overall end-to-end time for the complete active learning process. Train is the total time spent training the models across all rounds. Eval is the total time spent evaluating the model on the test set after each round. Query is the time spent in the acquisition step. This step includes

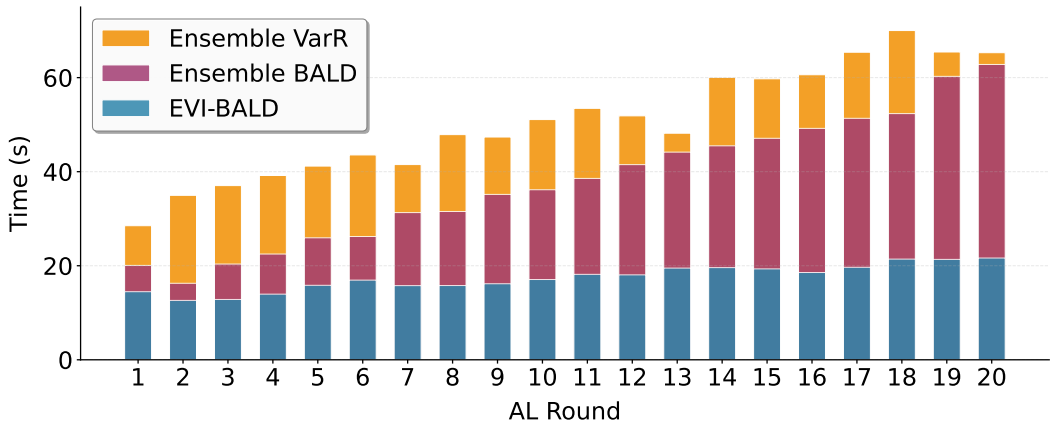

Figure 3: The bar chart illustrates the per-round computational time for each active learning experiment. Our proposed EVI-BALD method (blue) effectively reduces the computational cost of BALD by employing a hierarchical (evidential) model instead of Monte Carlo sampling.

scoring the unlabeled pool and selecting the next samples to label. Additionally, Per Round provides the average time per active learning round, computed by dividing the Total time by the number of rounds. This breakdown reveals the relative computational cost contributed by training, evaluation, and the acquisition step for each method.

For ensemble-based methods, we use an ensemble size of five models. This choice is reflected in Table 1, where the training time is approximately five times higher than that of our single-model Evi-BALD approach. In contrast, the MC dropout baseline exhibits relatively low computational cost, as it trains only one model with stochastic dropout rather than multiple independent networks. However, its active learning performance remains less stable due to the inherent stochasticity of the dropout process, as we can see in figure 1.

6 Conclusion and Limitation

The process of Monte Carlo sampling for BALD is computationally expensive. In this work, we propose *Evi-BALD*, a novel and efficient approximation of BALD that eliminates the need for ensemble-based Monte Carlo sampling. Instead of sampling multiple models, we train a single hierarchical model that directly parameterizes a second-order or evidential distribution over the first-order predictive distribution. This hierarchical structure provides a tractable approximation of the predictive posterior. To address the remaining challenge of intractable integration in the expected entropy term, we derive a closed-form second-order Taylor expansion of the entropy functional around the mean predictive distribution. This approximation reveals that BALD can be interpreted as capturing the sum of quadratic and higher-order remainder terms in the Taylor series of the entropy, enabling a computationally lightweight yet theoretically grounded reformulation.

We refer to our resulting acquisition function as Evi-BALD. Extensive experiments on CIFAR-10 demonstrate that Evi-BALD achieves leading performance compared to both classical BALD and other methods, while drastically reducing computational overhead by avoiding ensemble training and sampling. Our approach offers a scalable, single-model alternative for uncertainty-driven active learning in deep neural networks.

A key limitation of our approach lies in the second-order Taylor expansion approximation, where we neglect the higher-order remainder terms $\mathbb{E}_{\boldsymbol{\mu}}[R_3(\mathbb{P}, \mathbb{P}_0)]$ for computational tractability. This approximation is most accurate when the Dirichlet concentration parameter $\alpha_0 = \sum_{k=1}^{K} \alpha_k$ is large, as the posterior distribution $p(\boldsymbol{\mu} \mid \boldsymbol{\alpha}^*)$ becomes tightly concentrated around its mean $\boldsymbol{\mu}_0$, rendering higher-order terms negligible. However, the approximation may degrade when α_0 is small, indicating high epistemic uncertainty, the Dirichlet distribution becomes more diffuse and higher-order terms contribute more significantly to the entropy.

REFERENCES

Taiga Abe, Estefany Kelly Buchanan, Geoff Pleiss, Richard Zemel, and John P Cunningham. Deep ensembles work, but are they necessary? *Advances in Neural Information Processing Systems*, 35:33646–33660, 2022.

William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9368–9377, 2018.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. *Advances in Neural Information Processing Systems*, 35: 29205–29216, 2022.

Sylvain Calinon, Florent Guenter, and Aude Billard. On learning, representing, and generalizing a task in a humanoid robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(2):286–298, 2007.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in neural information processing systems*, 33:1356–1367, 2020.

Bertrand Charpentier, Oliver Borchert, Daniel Zügner, Simon Geisler, and Stephan Günnemann. Natural posterior network: Deep bayesian uncertainty for exponential family distributions. *arXiv preprint arXiv:2105.04471*, 2021.

Wenhu Chen, Yilin Shen, Hongxia Jin, and William Wang. A variational dirichlet framework for out-of-distribution detection. *arXiv preprint arXiv:1811.07308*, 2018.

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017.

Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). 2006.

Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pp. 417–424, 2006.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pp. 13–19. ACM New York, NY, USA, 1995.

Minghao Li, Junjie Qiu, and Weishi Shi. Label-wise uncertainty decomposition for multi-label classification by maximizing type ii likelihood.

David JC MacKay. The evidence framework applied to classification networks. *Neural computation*, 4(5):720–736, 1992.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

Andrey Malinin and Mark Gales. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. *Advances in neural information processing systems*, 32, 2019.

Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pp. 388–394. Springer, 1992.

Dan Roth and Kevin Small. Margin-based active learning for structured output spaces. In *European conference on machine learning*, pp. 413–424. Springer, 2006.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Dayou Yu, Minghao Li, Weishi Shi, and Qi Yu. Evidential mixture machines: Deciphering multi-label correlations for active learning sensitivity. *Advances in Neural Information Processing Systems*, 37:112217–112236, 2024.

A Entropy Decomposition for Bernoulli-Beta Model

According to equation 2, we have:

$$I(\mathbf{t}; \boldsymbol{\mu} \mid \mathbf{x}, \mathcal{D}) \approx H\left[\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[p(\mathbf{t} \mid \boldsymbol{\mu})\right]\right] - \mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[H(\mathbf{t} \mid \boldsymbol{\mu})\right].$$

For the left term, we have:

$$\begin{aligned}
&H\left[\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[p(\mathbf{t} \mid \boldsymbol{\mu})\right]\right] \\
&= H\Big[\mathbb{E}_{\mu\sim p(\mu|a^*,b^*,\mathbf{x},\mathcal{D})}\left[\mu^t(1-\mu)^{1-t}\right]\Big] \\
&= H\Big[\left(\mathbb{E}_\mu[\mu]\right)^t\left(\mathbb{E}_\mu[1-\mu]\right)^{1-t}\Big] \\
&= H\Big[\Big(\frac{a^*}{a^*+b^*}\Big)^t\Big(\frac{b^*}{a^*+b^*}\Big)^{1-t}\Big] \\
&= -\frac{a^*}{a^*+b^*}\log\frac{a^*}{a^*+b^*} - \frac{b^*}{a^*+b^*}\log\frac{b^*}{a^*+b^*}
\end{aligned}$$

where we substitute a^* and b^* for $\mathbf{m}^*$. Since $t \in \{0, 1\}$ is fixed, the expectation of the Bernoulli likelihood factorizes into powers of the posterior mean of μ and $1-\mu$. And for cases $t = 1$ and $t = 0$ are identical, due to symmetry.

According to equation 2, the second term in the Evi-BALD approximation is the expected binary entropy:

$$\mathbb{E}_{\mu\sim p(\mu|a^*,b^*)}\Big[H(p(\mathbf{t} \mid \mu))\Big] = \mathbb{E}_{\mu\sim p(\mu|a^*,b^*)}\Big[h(\mu)\Big],$$

where $h(\mu) = -\mu\log\mu - (1-\mu)\log(1-\mu)$ is the binary entropy function and $\mu \sim \text{Beta}(a^*, b^*)$.

Direct computation of this expectation is intractable in closed form. We therefore approximate it using the second-order Taylor expansion of $h(\mu)$ around the posterior mean $\mu_0 = \mathbb{E}[\mu] = \frac{a^*}{a^*+b^*}$:

$$\begin{aligned}
h(\mu) \approx\ & h(\mu_0) + h'(\mu_0)(\mu-\mu_0) \\
&+ \frac{1}{2}h''(\mu_0)(\mu-\mu_0)^2.
\end{aligned}$$

Taking the expectation over $\mu \sim p(\mu \mid a^*, b^*)$ yields

$$\begin{aligned}
\mathbb{E}_\mu\Big[h(\mu)\Big] &\approx h(\mu_0) + h'(\mu_0)\cdot\mathbb{E}[\mu-\mu_0] \\
&\quad + \frac{1}{2}h''(\mu_0)\cdot\mathbb{E}[(\mu-\mu_0)^2] \\
&= h(\mu_0) + \frac{1}{2}h''(\mu_0)\,\text{Var}(\mu),
\end{aligned}$$

since $\mathbb{E}[\mu-\mu_0] = 0$. And for simplicity, we will not specifically denote the remainders higher than the third order. We now compute each component explicitly. First, the binary entropy evaluated at the mean is

$$h(\mu_0) = -\frac{a^*}{a^*+b^*}\log\frac{a^*}{a^*+b^*} - \frac{b^*}{a^*+b^*}\log\frac{b^*}{a^*+b^*}.$$

The first derivative of the binary entropy is

$$h'(\mu) = \log\left(\frac{1-\mu}{\mu}\right),$$

but this term vanishes in expectation and is not needed for the second-order approximation.

The second derivative is

$$h''(\mu) = -\frac{1}{\mu(1-\mu)}.$$

Evaluating at $\mu_0 = \frac{a^*}{a^*+b^*}$ gives

$$h''(\mu_0) = -\frac{1}{\frac{a^*}{a^*+b^*} \cdot \frac{b^*}{a^*+b^*}} = -\frac{(a^*+b^*)^2}{a^*b^*}.$$

The variance of the Beta(a^*, b^*) distribution is

$$\text{Var}(\mu) = \frac{a^*b^*}{(a^*+b^*)^2(a^*+b^*+1)}.$$

Substituting these into the second-order term yields

$$\begin{aligned}
&\frac{1}{2}h''(\mu_0)\,\text{Var}(\mu) \\
&= \frac{1}{2}\left(-\frac{(a^*+b^*)^2}{a^*b^*}\right) \cdot \frac{a^*b^*}{(a^*+b^*)^2(a^*+b^*+1)} \\
&= -\frac{1}{2(a^*+b^*+1)}.
\end{aligned}$$

Therefore, the expected binary entropy is approximated as

$$\begin{aligned}
\mathbb{E}_{\mu\sim p(\mu|a^*,b^*)}\Big[h(\mu)\Big] \approx &-\frac{a^*}{a^*+b^*}\log\frac{a^*}{a^*+b^*} - \frac{b^*}{a^*+b^*}\log\frac{b^*}{a^*+b^*} \\
&- \frac{1}{2(a^*+b^*+1)}.
\end{aligned}$$

This second-order approximation becomes increasingly accurate when $a^* + b^*$ is large (i.e., when the Beta posterior is tightly concentrated around its mean).

B Entropy Decomposition for Multinomial-Dirichlet Model

We extend the approximation to the categorical case, where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K) \sim \text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_K)$ with $\alpha_0 = \sum_{j=1}^K \alpha_j$, and the multinomial entropy is $H(\boldsymbol{\mu}) = -\sum_{k=1}^K \boldsymbol{\mu}_k \log \boldsymbol{\mu}_k$. In the Categorical case, we will replace the second-order parameters $\mathbf{m}$ with $\boldsymbol{\alpha}$ in equation 2. The entropy of the expectation of the Categorical distribution is derived as follows:

$$\begin{aligned}
&H\left[\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\mathbf{m}^*,\mathbf{x},\mathcal{D})}\left[p(\mathbf{t}\mid\boldsymbol{\mu})\right]\right] \\
&= H\Big[\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\boldsymbol{\alpha}^*,\mathbf{x},\mathcal{D})}\Big[\prod_{k=1}^K \mu_k^{t_k}\Big]\Big] \\
&= H\Big[\prod_{k=1}^K \big(\mathbb{E}_{\boldsymbol{\mu}}[\mu_k]\big)^{t_k}\Big] \\
&= H\Big[\prod_{k=1}^K \Big(\frac{\alpha_k^*}{\sum_{j=1}^K \alpha_j^*}\Big)^{t_k}\Big] \\
&= -\sum_{k=1}^K \frac{\alpha_k^*}{\sum_{j=1}^K \alpha_j^*}\log\frac{\alpha_k^*}{\sum_{j=1}^K \alpha_j^*}
\end{aligned}$$

where we substitute $\boldsymbol{\alpha}^*$ for $\mathbf{m}^*$. Since $\mathbf{t} \in \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$ is a fixed one-hot vector (with $t_{k^*} = 1$ for some class k^* and 0 elsewhere), the expectation of the Categorical likelihood factorizes into powers of the posterior means of the μ_k.

According to equation 2, the second term in the Evi-BALD approximation for the categorical (multi-class) case is the expected multinomial entropy:

$$\mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\boldsymbol{\alpha}^*)}\Big[H(p(\mathbf{t}\mid\boldsymbol{\mu}))\Big] = \mathbb{E}_{\boldsymbol{\mu}\sim p(\boldsymbol{\mu}|\boldsymbol{\alpha}^*)}\Big[H(\boldsymbol{\mu})\Big],$$

where $H(\boldsymbol{\mu}) = -\sum_{k=1}^{K} \mu_k \log \mu_k$ is the multinomial entropy function and $\boldsymbol{\mu} \sim \text{Dirichlet}(\boldsymbol{\alpha}^*)$.

Direct computation of this expectation is intractable in closed form. We therefore approximate it using the second-order multivariate Taylor expansion of $H(\boldsymbol{\mu})$ around the posterior mean $\boldsymbol{\mu}_0 = \mathbb{E}[\boldsymbol{\mu}]$, where $(\boldsymbol{\mu}_0)_k = \mu_{0k} = \frac{\alpha_k^*}{\alpha_0^*}$ and $\alpha_0^* = \sum_{j=1}^{K} \alpha_j^*$:

$$\begin{aligned} H(\boldsymbol{\mu}) \approx H(\boldsymbol{\mu}_0) &+ \nabla H(\boldsymbol{\mu}_0)^\top (\boldsymbol{\mu} - \boldsymbol{\mu}_0) \\ &+ \frac{1}{2} (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \mathbf{Hess}\, H(\boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0). \end{aligned}$$

Taking the expectation over $\boldsymbol{\mu} \sim p(\boldsymbol{\mu} \mid \boldsymbol{\alpha}^*)$ yields

$$\begin{aligned} \mathbb{E}\Big[H(\boldsymbol{\mu})\Big] &\approx H(\boldsymbol{\mu}_0) + \nabla H(\boldsymbol{\mu}_0)^\top \cdot \mathbb{E}[\boldsymbol{\mu} - \boldsymbol{\mu}_0] \\ &\quad + \frac{1}{2} \mathbb{E}\Big[(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \mathbf{Hess}\, H(\boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\Big] \\ &= H(\boldsymbol{\mu}_0) + \frac{1}{2} \operatorname{tr}\Big(\mathbf{Hess}\, H(\boldsymbol{\mu}_0)\ \operatorname{Cov}(\boldsymbol{\mu})\Big), \end{aligned}$$

since $\mathbb{E}[\boldsymbol{\mu} - \boldsymbol{\mu}_0] = \mathbf{0}$.

We now compute each component explicitly.

The multinomial entropy evaluated at the mean is

$$H(\boldsymbol{\mu}_0) = -\sum_{k=1}^{K} \frac{\alpha_k^*}{\alpha_0^*} \log \frac{\alpha_k^*}{\alpha_0^*}.$$

The gradient of the entropy is

$$[\nabla H(\boldsymbol{\mu})]_j = -\log \mu_j - 1,$$

but this term vanishes in expectation and is not needed for the second-order approximation.

The Hessian matrix is diagonal (due to the separability of the entropy terms on the simplex):

$$[\mathbf{Hess}\, H(\boldsymbol{\mu})]_{jk} = -\frac{\delta_{jk}}{\mu_j},$$

so at the mean point

$$\mathbf{Hess}\, H(\boldsymbol{\mu}_0) = \operatorname{diag}\left(-\frac{1}{\mu_{01}}, -\frac{1}{\mu_{02}}, \ldots, -\frac{1}{\mu_{0K}}\right).$$

Because the Hessian is diagonal, the quadratic form simplifies to

$$\begin{aligned} &(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \mathbf{Hess}\, H(\boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0) \\ &= \sum_{k=1}^{K} -\frac{1}{\mu_{0k}} (\mu_k - \mu_{0k})^2. \end{aligned}$$

Its expectation is therefore

$$\begin{aligned} &\mathbb{E}\Big[(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^\top \mathbf{Hess}\, H(\boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\Big] \\ &= -\sum_{k=1}^{K} \frac{1}{\mu_{0k}} \operatorname{Var}(\mu_k). \end{aligned}$$

Under the Dirichlet distribution, the marginal distribution of each component is Beta:

$$\mu_k \sim \text{Beta}(\alpha_k^*, \alpha_0^* - \alpha_k^*),$$

so the marginal variance is

$$\mathrm{Var}(\mu_k) = \frac{\alpha_k^*(\alpha_0^* - \alpha_k^*)}{(\alpha_0^*)^2(\alpha_0^* + 1)} = \frac{\mu_{0k}(1 - \mu_{0k})}{\alpha_0^* + 1}.$$

Substituting gives

$$\frac{\mathrm{Var}(\mu_k)}{\mu_{0k}} = \frac{1 - \mu_{0k}}{\alpha_0^* + 1}.$$

Summing over all classes:

$$\begin{aligned}\sum_{k=1}^{K} \frac{\mathrm{Var}(\mu_k)}{\mu_{0k}} &= \frac{1}{\alpha_0^* + 1} \sum_{k=1}^{K} (1 - \mu_{0k}) \\ &= \frac{1}{\alpha_0^* + 1}(K - 1),\end{aligned}$$

since $\sum_{k=1}^{K} \mu_{0k} = 1$.

The second-order correction term is therefore

$$\begin{aligned}&\frac{1}{2}\left(-\sum_{k=1}^{K} \frac{\mathrm{Var}(\mu_k)}{\mu_{0k}}\right) \\ &= -\frac{1}{2} \cdot \frac{K - 1}{\alpha_0^* + 1} \\ &= -\frac{K - 1}{2(\alpha_0^* + 1)}.\end{aligned}$$

Thus, the expected multinomial entropy is approximated as

$$\mathbb{E}_{\boldsymbol{\mu} \sim p(\boldsymbol{\mu}|\boldsymbol{\alpha}^*)}\Big[H(\boldsymbol{\mu})\Big] \approx -\sum_{k=1}^{K} \frac{\alpha_k^*}{\alpha_0^*} \log \frac{\alpha_k^*}{\alpha_0^*} - \frac{K - 1}{2(\alpha_0^* + 1)}.$$

This second-order approximation becomes increasingly accurate when $\alpha_0^* = \sum_{j=1}^{K} \alpha_j^*$ is large, i.e., when the Dirichlet posterior is tightly concentrated around its mean $\boldsymbol{\mu}_0$.