Uncertainty Estimates of Predictions via a General Bias-Variance Decomposition

Sebastian G. Gruber
German Cancer Research Center (DKFZ)
German Cancer Consortium (DKTK)
Goethe University Frankfurt, Germany
`sebastian.gruber@dkfz.de`

Florian Buettner
German Cancer Research Center (DKFZ)
German Cancer Consortium (DKTK)
Frankfurt Cancer Institute, Germany
Goethe University Frankfurt, Germany
`florian.buettner@dkfz.de`

Abstract

Reliably estimating the uncertainty of a prediction throughout the model lifecycle is crucial in many safety-critical applications. The most common way to measure this uncertainty is via the predicted confidence. While this tends to work well for in-domain samples, these estimates are unreliable under domain drift and restricted to classification. Alternatively, proper scores can be used for most predictive tasks but a bias-variance decomposition for model uncertainty does not exist in the current literature. In this work we introduce a general bias-variance decomposition for proper scores, giving rise to the Bregman Information as the variance term. We discover how exponential families and the classification log-likelihood are special cases and provide novel formulations. Surprisingly, we can express the classification case purely in the logit space. We showcase the practical relevance of this decomposition on several downstream tasks, including model ensembles and confidence regions. Further, we demonstrate how different approximations of the instance-level Bregman Information allow reliable out-of-distribution detection for all degrees of domain drift.

1 INTRODUCTION

A core principle behind the success of modern Machine and Deep Learning approaches are loss functions, which are used to optimize and compare the goodness-of-fit of predictive models. Typical loss functions, such as the

Brier score or the negative log-likelihood, capture not only predictive power (in the sense of accuracy) but also predictive uncertainty. The latter is particularly relevant in sensitive forecasting domains, such as cancer diagnostics (Haggenmüller et al., 2021), genotype-based disease prediction (Katsaouni et al., 2021) or climate prediction (Yen et al., 2019).

Proper scores are a common occurrence as loss functions for probabilistic modelling since their defining criterion is to assign the best value to the target distribution as prediction. Consequently, they are widely applicable from quantile regression (Gneiting & Raftery, 2007) to generative models (Song et al., 2021). They are a generalization of the log-likelihood and also cover exponential families (Grünwald & Dawid, 2004). However, for such loss functions, it is not clear how we can decompose them such that a component capturing predictive uncertainty arises. Consequently, predictive uncertainty is typically only considered as variance of predictions or, in classification, via the confidence score associated to the top-label prediction. Such confidence scores capture the predictive uncertainty well if they are calibrated, namely if the confidence of a prediction matches its true likelihood (Guo et al., 2017). However, the calibration error of these confidence scores typically increases under domain drift, making them an unreliable measure for predictive uncertainty in many real-world applications (Ovadia et al., 2019; Tomani & Buettner, 2021).

In this work, we discover the Bregman Information as a natural replacement of model variance via a bias-variance decomposition for proper scores. The Bregman Information generalizes the variance of a random variable via a closed-form definition based on a generating function (Banerjee et al., 2005). In the case of our decomposition, this generating function is a convex conjugate directly associated with the respective proper score. The source code for the experiments is openly accessible at `https://github.com/MLO-lab/Uncertainty_Estimates_via_BVD`. We summarize our **contributions** in the following:

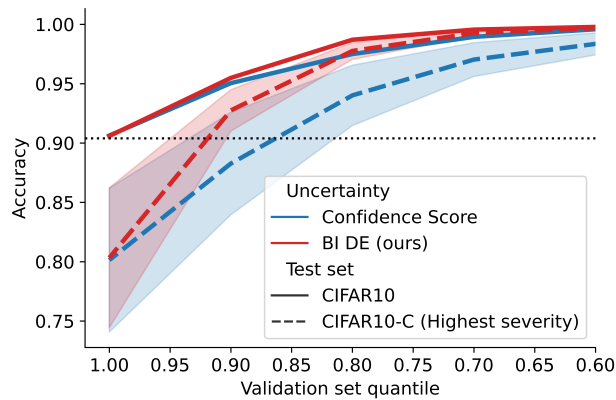- In Section 3, we generalize relevant properties to func-

Figure 1: Accuracy after discarding test instances with high levels of uncertainty. We can discard fewer samples to reach better accuracy when using Bregman Information as uncertainty measure. For example, to achieve the validation set accuracy (dotted line) for severely corrupted data we only have to discard ~7% of most uncertain in-domain samples contrary to ~14% when using the confidence score. The standard deviation bounds stem from different types of corruption.

tional Bregman divergences, which allows for deriving a bias-variance decomposition for strictly proper scores. Via Bregman Information, we give novel formulations for decompositions of exponential families and the classification log-likelihood in the logit space.

- We generalize the law of total variance to Bregman Information and show how ensemble predictions marginalize out a specific source of uncertainty in Section 3.5. We also propose a general way to give confidence regions for predictions in Section 3.6.

- We showcase experiments on how typical classifiers differ in their Bregman Information in Section 4. There, we demonstrate that the Bregman Information can be a more meaningful measure of out-of-domain uncertainty compared to the confidence score in the case of corrupted CIFAR-10 and ImageNet (Figure 1 and Algorithm 1).

2   BACKGROUND

In this section, we first start with a basic introduction of Bregman divergences and Bregman Information. We specifically mention recent developments for functional Bregman divergences as we will require and provide generalizations to this topic. Then follows another introduction into the basic concepts of proper scores and exponential families, which are related to Bregman divergences. Finally, we will discuss other proposed bias-variance decompositions in the literature to put our contribution into perspective.

2.1   Bregman Divergences and Bregman Information

Bregman divergences are a class of divergences occurring in a wide range of applications (Bregman, 1967; Banerjee et al., 2005; Frigyik et al., 2006; Si et al., 2009; Gupta et al., 2022). We use the following definition.

Definition 2.1 (Bregman (1967); Banerjee et al. (2005))**.** Given an open convex set $U \subset \mathbb{R}^d$, let $\phi: U \to \mathbb{R}$ be a differentiable, convex function. The **Bregman divergence** generated by ϕ is defined as

$$d_\phi(x, y) = \phi(y) - \phi(x) - \langle \nabla\phi(x), y - x \rangle.$$

It can be interpreted geometrically as the difference between ϕ and the supporting tangent plane of $\phi(x)$ at y. We have $d_\phi(x, y) \geq 0$ with $d_\phi(x, y) = 0$ if $x = y$.

By definition, we can use Bregman divergences for scalar and vector inputs. But, in the infinite-dimensional case, for example when dealing with a continuous distribution space $\mathscr{P}$, the gradient vector and the inner product are not defined anymore. As a solution to this, Frigyik et al. (2006) introduce functional Bregman divergences by replacing the inner product term with the Fréchet derivative. The authors showed that the functional case generalizes the standard case. Since some relevant functions are not Fréchet differentiable, Ovcharov (2018) offers an alternative approach to define the functional case for a set of distributions $\mathscr{P}$. They propose to use the space $\mathscr{L}(\mathscr{P})$ of $\mathscr{P}$-integrable functions and the space $\mathrm{span}\mathscr{P}$ of finite linear combinations of elements from $\mathscr{P}$. Note that these two spaces are dual with the pairing "·" defined as $f \cdot P = \int f \mathrm{d}P$ for $f \in \mathscr{L}(\mathscr{P})$ and $P \in \mathrm{span}\mathscr{P}$. This operator is similar to the inner product for categorical distributions. Further, they use the definition of subgradients: A subgradient x' at point $x \in U \subset \mathrm{span}\mathscr{P}$ of a function $\phi: U \to \mathbb{R}$ fulfills the property $\phi(y) \geq \phi(x) + x' \cdot (y - x)$ for all $y \in U$. A function ϕ' which maps to a subgradient of $\phi(x)$ for all $x \in U$ is called a selection of subgradients, or, if it is unambiguous in the context, just subgradient of ϕ. In general, subgradients are not unique, unlike gradients. Ovcharov (2018) proceeds to define the by (ϕ, ϕ') generated **functional Bregman divergence** as $d_{\phi, \phi'}(x, y) = \phi(y) - \phi(x) - \phi'(x) \cdot (y - x)$. We will use this formulation for a general bias-variance decomposition of proper scores, which we introduce in Section 3. The **convex conjugate** ϕ^* of a function ϕ is defined as $\phi^*(x^*) = \sup_y \langle x^*, y \rangle - \phi(y)$ (Rockafellar, 1970). If ϕ is differentiable and strictly convex, then $(\nabla\phi)^{-1} = \nabla\phi^*$ and $\phi^{**} = \phi$. For this case, Banerjee et al. (2005) give the important fact that $d_\phi(x, y) = d_{\phi^*}(\nabla\phi(y), \nabla\phi(x))$. That is, by using the convex conjugate, we can flip the arguments in a Bregman divergence. To derive our main contribution,
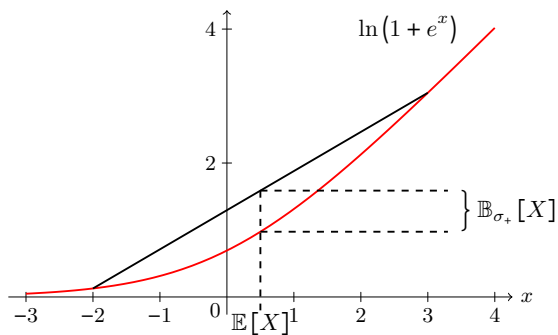
Figure 2: Illustration of the Bregman Information generated by the softplus function $\sigma_+(x) = \ln(1 + e^x)$ of a binary random variable.

we will state a generalization of this property to functional Bregman divergences in Lemma 3.1.

We can also use Bregman divergences to quantify the variability or deviation of a random variable. Throughout this work, the following definition is a central concept.

Definition 2.2 (Banerjee et al. (2005)). Given a convex set U, let $\phi: U \to \mathbb{R}$ be a differentiable, convex function. The **Bregman Information** (generated by ϕ) of a random variable X with realizations in U is defined as

$$\mathbb{B}_\phi[X] = \mathbb{E}[d_\phi(\mathbb{E}[X], X)].$$

The Bregman Information generalizes the variance of a random variable since both are equal if we set $U = \mathbb{R}$ and $\phi(x) = x^2$. Thus, one interpretation of the Bregman Information is that it measures the divergence of a random variable from its mean. Another representation, which does not depend on d_ϕ, is $\mathbb{B}_\phi[X] = \mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X])$. Banerjee et al. (2005) show that this follows from the original definition. Recall that Jensen's inequality gives $\mathbb{E}[\phi(X)] \geq \phi(\mathbb{E}[X])$. Consequently, a second interpretation of the Bregman Information is that it measures the gap between both sides of Jensen's inequality of the convex function ϕ and random variable X (Banerjee et al., 2005). It also shows that a generalization to the functional case is invariant towards the chosen subgradient of the underlying functional Bregman divergence. Thus, we omit the subgradient when generating functional Bregman Information. The Bregman Information generated by the *softplus* function is depicted in Figure 2 for a binary random variable. The softplus finds use as an activation function in neural networks (Glorot et al., 2011; Murphy, 2022). Its generalization is the so-called LogSumExp function (c.f. Section 3.4). In Section 3, the Bregman Information will play a critical role in our bias-variance decompositions since it represents

the variance term. Further, the LogSumExp-generated version covers the variance term for classification. It reduces to the softplus version for the binary case.

2.2 Proper Scores and Exponential Families

Gneiting & Raftery (2007) give an extensive and approachable overview of proper scores. In short, proper scores put a negative loss on a distribution prediction $P \in \mathscr{P}$ for a target observation $Y \sim Q \in \mathscr{P}$ and reach their maximum if $P = Q$. For a concise statement of our main result, we require a more technical definition provided in the following similar to Hendrickson & Buehler (1971), Ovcharov (2015), and Ovcharov (2018). We call a function $S: \mathscr{P} \to \mathscr{L}(\mathscr{P})$ **scoring rule** or just **score**. Note that for a given P, $S(P)$ maps into a function space and can be again evaluated on an observation y, like $S(P)(y)$. To assess the goodness-of-fit between distributions P and Q, we use the expected score $S(P) \cdot Q = \mathbb{E}_{Y \sim Q}[S(P)(Y)]$. A score is defined to be **proper** on $\mathscr{P}$ if and only if $S(P) \cdot Q \leq S(Q) \cdot Q$ holds for all $P, Q \in \mathscr{P}$, and **strictly proper** if and only if an equality implies $P = Q$. In other words, a score is proper if predicting the target distribution gives the best expectation and strictly proper if no other prediction can achieve this value. Note that the choice of $\mathscr{P}$ is relevant: The negative squared error of a mean prediction is strictly proper for normal distributions with fixed variance but only proper if the variance varies. Given a proper score, the associated **negative entropy** $G: \mathscr{P} \to \mathbb{R}$ is defined as $G(Q) = S(Q) \cdot Q$. It represents the highest reachable value for a given target. Further, it has S as a subgradient and is (strictly) convex if and only if S is (strictly) proper. Ovcharov (2018) proved that a proper score is closely related to a functional Bregman divergence generated by the associated negative entropy via $G(Q) - S(P) \cdot Q = d_{G,S}(P, Q)$. An example of such a relation is the Kullback-Leibler divergence and the Shannon entropy associated with the log score (log-likelihood).

Next, we summarize relevant aspects of exponential families. Banerjee et al. (2005) provides a more extensive introduction. For a support set $\mathscr{T}$, the probability density/mass function p_θ at a point $x \in \mathscr{T}$ of an **exponential family** is given by $p_\theta(x) = \exp(\langle \theta, T(x) \rangle - A(\theta))h(x)$. Here, we call $\theta \in \Theta$ the natural parameter of the convex parameter space $\Theta \subset \mathbb{R}^d$, T is the sufficient statistic, and A is the log-partition. Table 1 gives two relevant examples and shows the mapping between typical and natural parameters. Further examples are the Dirichlet, exponential, and Poisson distributions. There are two relevant properties which we will require for our results. One is that A is a strictly convex function. The other is $\mathbb{E}[T(X)] = \nabla A(\theta)$ for $X \sim p_\theta$. Banerjee et al. (2005) proved under mild conditions that an exponential family relates to a Bregman divergence and vice versa via the negative log-likelihood. Grünwald & Dawid (2004) proved a similar link between proper scores and exponential families.

Table 1: Examples of exponential families. The mapping defines the relation between natural parameters and typical parameters. We denote the dummy-encoding for a $x \in \{1, \dots, k\}$ with d_x, class probabilities with p_j, mean with μ, and standard deviation with σ. The Bernoulli distribution is a special case of the categorical distribution for $k = 2$.

Distribution	$\mathscr{T}$	Θ	$T(x)$	$A(\theta)$	$h(x)$	Mapping
Categorical (k-classes)	$\{1, \dots, k\}$	$\mathbb{R}^{k-1}$	d_x	$\ln\left(1 + \sum_{i=1}^{k-1} \exp\theta_i\right)$	1	$p_j = \frac{\exp\theta_j}{1+\sum_{i=1}^{k-1}\exp\theta_i}$
Normal (known σ)	$\mathbb{R}$	$\mathbb{R}$	x	$\frac{\theta^2}{2}$	$\frac{\exp\left(\frac{-x^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma}$	$\mu = \theta\sigma$

As we can see, exponential families, proper scores, and Bregman divergences have strong relationships to one another. Unfortunately, the literature addresses most cases and proves only for vector-based Bregman divergences. One contribution of this work is to generalize the necessary properties to the functional case. Then, it possible to derive a general bias-variance decomposition of proper scores, where we will discover a functional Bregman Information as representative of the variance term. In the case of the log-likelihood of an exponential family, the decomposition reduces to a vector-based Bregman Information.

2.3 Other Bias-Variance Decompositions

In general, all decompositions in current literature are either for categorical, real-valued, or parametric predictions, and it is not clear if a decomposition for proper scores of non-parametric distributions is possible.

James & Hastie (1997) formulate a decomposition for any loss function of categorical or real-valued predictions but do not provide a closed-form solution for a given case. Domingos (2000) introduce how a general bias-variance decomposition should look, though they stated it is unclear when or if this decomposition holds for a loss function. James (2003) provide a bias-variance decomposition for symmetric loss functions. Heskes (1998) use the bias-variance decompositions for the Kullback-Leibler divergence, which allows to derive a decomposition for exponential families. Hansen & Heskes (2000) proves that a bias-variance decomposition of a parametric prediction is only possible if the prediction belongs to an exponential family. Importantly, they only introduce the specific decomposition for a given exponential family. The decomposition is not formulated for the natural parameters and relies on the canonical link function. Consequently, a relation to Bregman divergences and Bregman Information is missing, which we will provide.

A Pythagorean-like theorem for vector-based Bregman divergences is a known fact in literature (Jones & Byrne, 1990; Csiszar, 1991; Della Pietra et al., 2002; Dawid, 2007; Telgarsky & Dasgupta, 2012). An equality in this theorem implies a decomposition in the form of $\mathbb{E}\left[d_\phi(X, y)\right] = \mathbb{E}\left[d_\phi(X, x^*)\right] + d_\phi(x^*, y)$ with $x^* = \arg\min_z \mathbb{E}\left[d_\phi(X, z)\right]$ (Pfau, 2013). Brofos et al. (2019), Brinda et al. (2019), and Yang et al. (2020) relate the

classification log-likelihood to the Kullback-Leibler divergence and provide a bias-variance decomposition, where X takes the form of a predictive probability vector. They set $x^* \propto \exp\mathbb{E}\left[\log X\right]$. Note that predictions in the logit space require normalization to the log space, which will not be the case in our formulation. Gupta et al. (2022) build upon the Bregman divergence decomposition and use the notion of primal and dual space of the variance. Even though the definitions are similar, the authors did not state the relation between Bregman Information and dual variance, for which they introduce a general law of total variance.

Due to the restriction of Bregman divergences to vector inputs, it is not clear if a decomposition for proper scores of non-parametric distributions is possible. In other words, we require an extension of the current literature to functional Bregman divergences for a positive result. In the following section, we provide the required generalization and unify the variance term in previous literature via the Bregman Information.

3 A GENERAL BIAS-VARIANCE DECOMPOSITION

In this section, we offer a general bias-variance decomposition for proper scores. The only assumption is that the score is strictly proper for a distribution set $U \subset \mathscr{P}$ and that each respective expectation is finite. Further, we will discover that the variance term is the Bregman Information generated by the convex conjugate of the associated negative entropy. This discovery generalizes and unifies decompositions in current literature for which exists a concrete form (Hansen & Heskes, 2000), but also provides a closed formulation contrary to other general bias-variance decompositions (James & Hastie, 1997).

3.1 Functional Bregman Divergences of Convex Conjugates

The essential part for deriving our main result is the exchange of arguments in functional Bregman divergences. Note that a subgradient of a strictly convex function is injective. Thus, its inverse exists on an appropriate domain, and this inverse is again a subgradient of the convex conjugate.

With that in mind, we can state the following.

Lemma 3.1. *For strictly convex* $G: U \to \mathbb{R}$ *with subgradient* S, *the divergences* $d_{G,S}$ *and* $d_{G^*,S^{-1}}$ *are functional Bregman divergences with the properties*

- $d_{G,S}(p, q) = d_{G^*,S^{-1}}(S(q), S(p))$, *and*

- $d_{G^*,S^{-1}}(p^*, q^*) = d_{G,S}(S^{-1}(q^*), S^{-1}(p^*))$.

Further, given a random variable Q^* *we have*

$$\mathbb{E}\left[d_{G^*,S^{-1}}(p^*, Q^*)\right] = d_{G^*,S^{-1}}(p^*, \mathbb{E}[Q^*]) + \mathbb{B}_{G^*}[Q^*].$$

The last property is a generalization of the decomposition of Bregman divergences and uses Definition 2.2. Note that we can only define $d_{G^*,S^{-1}}$ on the range of S for its first argument. The second argument can be any linear combination of subgradients at points of G. We present the exact technical details with the proof in Appendix B.

This lemma confirms that a critical property of Bregman divergences is also well-defined for functional Bregman divergences. Namely, we can exchange the arguments by changing the generating convex function to the convex conjugate in a dual space. This insight leads us now to the main theoretical contribution of this work.

3.2 A Decomposition for Proper Scores

In the following, we present a bias-variance decomposition for strictly proper scores. Surprisingly, we will not have to make any additional assumptions, neither about the score or its entropy being differentiable nor about the set of predictions being convex.

Theorem 3.2. *For a strictly proper score* S *with associated negative entropy* G, *an estimated prediction* $\hat{f}$, *and the target* $Y \sim Q$, *we have*

$$\underbrace{\mathbb{E}\left[-S(\hat{f})(Y)\right]}_{\text{Error}} =$$
$$\underbrace{-G(Q)}_{\text{Noise}} + \underbrace{\mathbb{B}_{G^*}\left[S(\hat{f})\right]}_{\text{"Variance"}} + \underbrace{d_{G^*,S^{-1}}\left(S(Q), \mathbb{E}\left[S(\hat{f})\right]\right)}_{\text{Bias}}.$$

As we can see, the variance term is always represented by the Bregman Information generated by the convex conjugate of the associated negative entropy. The theorem directly follows from the relationship between proper scores and functional Bregman divergences provided by Ovcharov (2018) in combination with Lemma 3.1 (c.f. Appendix B). We provide an example in the form of the prominent log score. For conciseness, we denote all distributions with their respective densities.

Example 3.3. The log score $S = \ln$ is strictly proper on any set of densities. Its negative entropy is the negative Shannon entropy $H(p) = \int p(x) \ln p(x) \, dx$. The convex conjugate is the log partition function $H^*(p^*) = \ln \int e^{p^*(x)} dx$. Since $\mathbb{E}\left[H^*\left(\ln(\hat{f})\right)\right] = 0$, we receive the Bregman Information in the form $\mathbb{B}_{H^*}\left[S(\hat{f})\right] = -H^*\left(\mathbb{E}\left[\ln(\hat{f})\right]\right)$.

Dealing with non-parametric distributions can be challenging both in theory and practice. Consequently, we provide the following restriction to exponential families and then neatly express the decomposition through the natural parameters. Specifically, $\mathbb{B}_{H^*}$ will be reduced to a vector-based Bregman Information.

3.3 Exponential Families as a Special Case

When we deal with densities, it is often more practical to consider parametric densities since they can be represented by a parameter vector instead of a function. Particularly relevant classes are exponential families for which one uses the log-likelihood to assess the goodness of fit. In the last section, we derived the decomposition for the log-likelihood expressed in the function space in Example 3.3. Thus, we provide the following special case as a novel formulation of (Hansen & Heskes, 2000).

Proposition 3.4. *For an exponential family density/mass function* $p_{\hat{\theta}}$ *with estimated natural parameter vector* $\hat{\theta}$, *log-partition* A, *and reference function* h, *we have the log likelihood decomposition*

$$\underbrace{\mathbb{E}\left[-\ln p_{\hat{\theta}}(Y)\right]}_{\text{NLL}} = \underbrace{n(\theta)}_{\text{Noise}} + \underbrace{\mathbb{B}_A\left[\hat{\theta}\right]}_{\text{"Variance"}} + \underbrace{d_A\left(\theta, \mathbb{E}\left[\hat{\theta}\right]\right)}_{\text{Bias}}$$

where $n(\theta) = -A^*(\nabla A(\theta)) - \mathbb{E}[\ln h(Y)]$ *and* $Y \sim p_\theta$.

The proof in Appendix B uses $\theta = \nabla A^*(\mathbb{E}[T(Y)])$ in the last step, where T is the sufficient statistic. This is a fundamental property of exponential families but only holds if the distribution assumption holds with the data. Since this is usually not the case in practical scenarios, it is important to note that the decomposition still holds if we replace every θ with $\nabla A^*(\mathbb{E}[T(Y)])$.

Example 3.5. We can recover the textbook decomposition of the MSE by using $Y \sim \mathcal{N}(\mu, \sigma^2)$ with unknown mean μ and known variance σ^2. The necessary information is provided in Table 1. For the LHS in Proposition 3.4 we have $\mathbb{E}\left[-\ln p_{\hat{\theta}}(Y)\right] = \frac{\mathbb{E}[(Y-\hat{\mu})^2]}{2\sigma^2} + \ln \sqrt{2\pi}\sigma$. And for the RHS, we get $n(\theta) = \frac{1}{2} + \ln \sqrt{2\pi}\sigma$, $\mathbb{B}_A\left[\hat{\theta}\right] = \frac{1}{2\sigma^2}\mathbb{V}[\hat{\mu}]$, and $d_A\left(\theta, \mathbb{E}\left[\hat{\theta}\right]\right) = \frac{(\mu-\mathbb{E}[\hat{\mu}])^2}{2\sigma^2}$. Finally, we subtract $\ln \sqrt{2\pi}\sigma$ on both sides and then multiply with $2\sigma^2$, resulting in

$$\mathbb{E}\left[(Y - \hat{\mu})^2\right] = \sigma^2 + \mathbb{V}[\hat{\mu}] + (\mu - \mathbb{E}[\hat{\mu}])^2.$$

3.4 Classification Log-Likelihood as a Special Special Case

Even though classification is a particularly relevant task for Deep Learning, the current literature states the log-likelihood decomposition via the log-probabilities (Brofos et al., 2019; Brinda et al., 2019; Yang et al., 2020). Since neural networks output logits, a normalization step is required. This step is cumbersome to compute and hinders theoretical analysis. In the following, we will construct the Bregman Information for classification via Proposition 3.4 similar to Example 3.5. Surprisingly, the Bregman Information measures the variability of the prediction in the logit space and does not require normalization to the log-probability space.

Corollary 3.6. *For logit prediction $\hat{z} \in \mathbb{R}^k$ and target $Y \sim Q$ with k classes, we have*

$$\underbrace{\mathbb{E}\left[-\ln \mathrm{sm}_Y(\hat{z})\right]}_{\text{Classif. NLL}} = \underbrace{H(Q)}_{\text{Noise}} + \underbrace{\mathbb{B}_{\mathrm{LSE}}\left[\hat{z}\right]}_{\text{"Variance"}} + \underbrace{d_{\mathrm{LSE}}\left(\mathrm{sm}^{-1}(Q), \mathbb{E}\left[\hat{z}\right]\right)}_{\text{Bias}}$$

with the LogSumExp function $\mathrm{LSE}(x_1, \ldots, x_n) = \ln \sum_{i=1}^n e^{x_i}$, *the softmax function* $\mathrm{sm} = \nabla\mathrm{LSE}$, *and Shannon entropy H.*

The proof in Appendix B combines Proposition 3.4 with the properties for categorical distributions presented in Table 1. Note that sm^{-1} maps only into a $k-1$ dimensional subspace of $\mathbb{R}^k$.

The Bregman Information acting in the logit space is a convenient surprise for Deep Learning applications. Almost all neural networks used for classification do not output probabilities but logits. Consequently, we do not require normalizing the neural network outputs to compute the Bregman Information.

Further, we can express the Bregman Information in the binary classification case through the softplus function since $\mathbb{B}_{\mathrm{LSE}}\left[(\hat{z}_1, \hat{z}_2)^\mathsf{T}\right] = \mathbb{B}_{\sigma_+}\left[\hat{z}_2 - \hat{z}_1\right]$. The Bregman Information generated by the softplus function is illustrated in Figure 2. We chose the depicted random variable such that the Jensen gap visualizes geometrically.

3.5 Ensemble Predictions

Using ensembles as predictions in Machine Learning is a central aspect of many successful architectures, like Random Forest (Breiman, 2001), XGBoost (Chen & Guestrin, 2016), Deep Ensembles (DE) (Lakshminarayanan et al., 2017), or Test-Time augmentation (TTA) (Wang et al., 2019). DE and TTA show strong empirical results even though they do not sample the training data, unlike Random Forests or XGBoost. In the following, we show that an ensemble reduces a specific source of uncertainty via a general law of total variance. The presented contributions are generalizations of known concepts in the ensemble literature (Ueda & Nakano, 1996; Gupta et al., 2022). First, we require the following definition.

Definition 3.7. Given a convex set U, let $\phi\colon U \to \mathbb{R}$ be a convex function. We define the **conditional Bregman Information** (generated by ϕ) of a random variable X with realizations in U given another random variable Y as

$$\mathbb{B}_\phi\left[X \mid Y\right] = \mathbb{E}\left[d_\phi\left(\mathbb{E}\left[X \mid Y\right], X\right) \mid Y\right].$$

Similar to Bregman Information, it holds that $\mathbb{B}_\phi\left[X \mid Y\right] = \mathbb{E}\left[\phi(X) \mid Y\right] - \phi\left(\mathbb{E}\left[X \mid Y\right]\right)$. Again, the subgradient does not influence the functional case. The conditional variance appears by setting $\phi(x) = x^2$. We can now contribute the following.

Proposition 3.8 (Properties of Bregman Information)**.** *The general law of total variance for a (functional) Bregman Information $\mathbb{B}_G$ and random variables X and Y is given by*

$$\mathbb{B}_G\left[X\right] = \mathbb{E}\left[\mathbb{B}_G\left[X \mid Y\right]\right] + \mathbb{B}_G\left[\mathbb{E}\left[X \mid Y\right]\right]. \quad (1)$$

For i.i.d. random variables $X_1, \ldots, X_{2^n}$ and strictly convex G, we have

$$\mathbb{B}_G\left[\frac{1}{2^n} \sum_{i=1}^{2^n} X_i\right] < \mathbb{B}_G\left[\frac{1}{2^{n-1}} \sum_{i=1}^{2^{n-1}} X_i\right], \quad (2)$$

and if $n \to \infty$ and G is also continuous then

$$\mathbb{B}_G\left[\frac{1}{n} \sum_{i=1}^n X_i\right] \xrightarrow{a.s.} 0. \quad (3)$$

The last two properties also hold in the case of conditional Bregman Information. We now apply Corollary 3.8 in the context of ensemble predictions. To simplify the argument, we assume the context of an exponential family. But, the statements also hold for non-parametric scenarios. Let a prediction $\theta_{W,D}$ in the natural parameter space depend on W and D as two sources of variability. In the case of Deep Ensembles, D is the training data, and W is the weight initialization. For TTA, W is the input variation, like angle or rotation. If we compute an ensemble prediction $\hat{\theta}_D^{(n)} = 2^{-n} \sum_{i=1}^{2^n} \theta_{W_i,D}$ by sampling W while keeping D fixed, we marginalize out the uncertainty in the prediction stemming from W, since if $n \to \infty$, then

$$\mathbb{B}_A\left[\hat{\theta}_D^{(n)}\right] = \mathbb{B}_A\left[\mathbb{E}\left[\theta_{W,D} \mid D\right]\right] + \underbrace{\mathbb{E}\left[\mathbb{B}_A\left[\hat{\theta}_D^{(n)} \mid D\right]\right]}_{\to 0}. \quad (4)$$

As n does not influence the bias term, the expected negative log-likelihood reduces by the amount of conditional Bregman Information of W. Together with Equation (2), this results in

$$\mathbb{E}\left[-\ln p_{\hat{\theta}_D^{(n)}}(Y)\right] < \mathbb{E}\left[-\ln p_{\hat{\theta}_D^{(n-1)}}(Y)\right]. \quad (5)$$
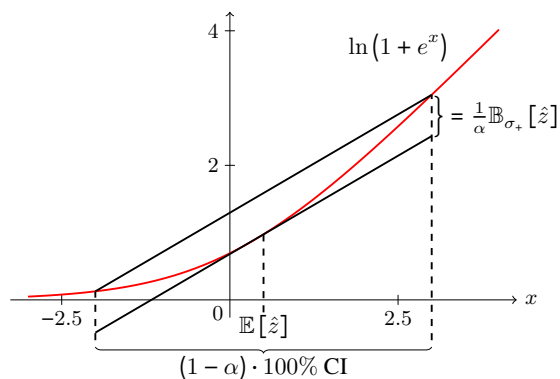
Figure 3: Illustration of the confidence interval for binary classification. The CI covers the area where the shifted tangent at $\mathbb{E}[\hat{z}]$ is larger than the generating convex function. This illustration generalizes to higher dimensions, where the CI is not an interval anymore in the strict sense but a region.

For $n = 0$ we recover the case of a single prediction. Equation (5) states that an ensemble always has better expected performance than a single model as long as the ensemble members come from a true source of uncertainty.

3.6 Confidence Regions Based on Bregman Information

In practical applications, one might be interested in a confidence interval for a prediction. For example, we could ask for an interval that covers a given prediction in 95% of the cases concerning the randomness in the training data. If we have a high-dimensional or functional prediction, we are not using an interval anymore but a convex set. Consequently, we refer to it as a confidence region. We can construct such a region as in the following. Applying Markov's inequality to the Bregman divergence d_G between a mean and its random variable $\mathscr{X}$ gives $\mathbb{P}(d_G(\mathbb{E}[X], X) \geq c) \leq \frac{1}{c}\mathbb{B}_G[X]$. Setting $c = \frac{1}{\alpha}\mathbb{B}_G[X]$ results in $d_G(\mathbb{E}[X], X) \leq \frac{1}{\alpha}\mathbb{B}_G[X]$ having at least probability $(1 - \alpha)$. Consequently, we are given a $(1 - \alpha) \cdot 100\%$-confidence region by

$$\mathrm{CR}_{(1-\alpha)} = \left\{ x \in \mathscr{X} \mid d_G(\mathbb{E}[X], x) \leq \frac{\mathbb{B}_G[X]}{\alpha} \right\} \quad (6)$$

with support $\mathscr{X}$ of X. An illustration for the binary classification case is depicted in Figure 3, where we construct a confidence interval in the one-dimensional logit space. Further, we demonstrate the confidence regions of a neural network trained on the Iris dataset. We estimate the Bregman Information by training models with different weight initializations. The resulting regions in Figure 4 illustrate the influence of the weight initialization on the variability of the prediction.
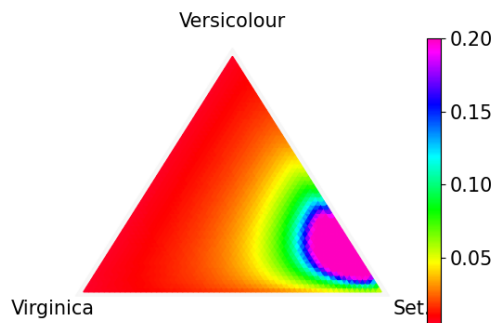

Figure 4: Confidence regions of a prediction transformed into the simplex for various alphas. The model is a simple neural network fitted on the Iris dataset.

4 EXPERIMENTS

In this section, we evaluate the Bregman Information and its approximations for classifiers via various experiments. Throughout all evaluations, we use the estimator $\hat{\mathbb{B}}_{\mathrm{LSE}}^{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{LSE}(\hat{z}_i) - \mathrm{LSE}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{z}_i\right)$. First, we evaluate typical classifiers on toy tasks, where we can simulate the ground truth. Accessing the data generation process also allows us to compare different approximation schemes of the model Bregman Information. Based on the insights, we provide experiments on corrupted CIFAR-10 and ImageNet and investigate how we may use the estimated Bregman Information for better predictive performance under domain drift.

4.1 Bregman Information of Common Classifiers

We assess the Bregman Information of the classifiers k-nearest neighbors, Support Vector Machine, Gaussian Process, Decision Tree, Random Forest, XGBoost, Naive Bayes, and neural network (Murphy, 2022). We create toy tasks and sample an arbitrary number of training datasets. For each sampled dataset, we fit all of the previous classifiers. This way, we can approximate the ground truth Bregman Information of each classifier arbitrarily well for growing sample size. Empirically, we consider 64 samples as a sufficiently close approximation. Further details appear in Appendix C. We plot the Bregman Information in a close region around the data distribution in the second row of Figure 5. The first row depicts a single exemplary sample of a training set and the confidence score of each respective classifier to put the Bregman Information plots in a meaningful perspective. As we can see, most models show a high uncertainty along the decision boundary. For example, the uncertainty of SVMs and Gaussian Processes is restricted to an area close to the training distribution. The Bregman Information of KNN and decision-tree-based models suggests
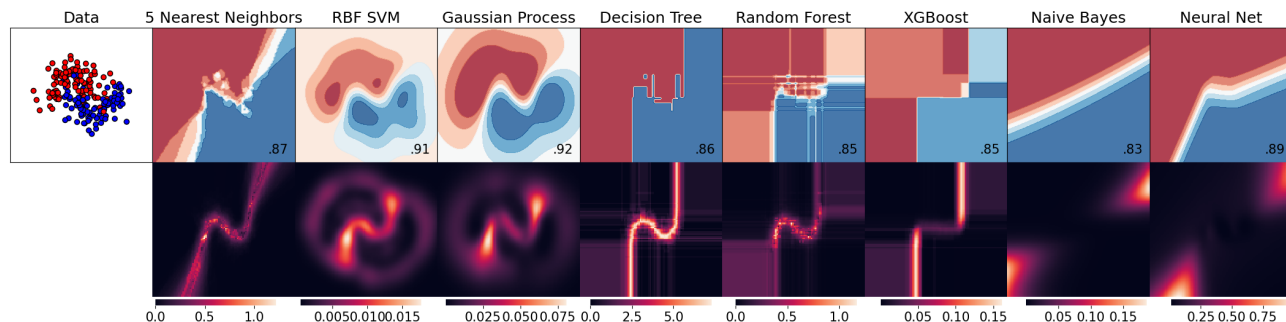
Figure 5: Several classifiers are trained on a simulated toy task. **Top**: Classifier predictions for a frame of the input space. The accuracy is displayed in each bottom right corner. **Bottom**: Bregman Information of these classifiers in the identical space based on multiple training runs.
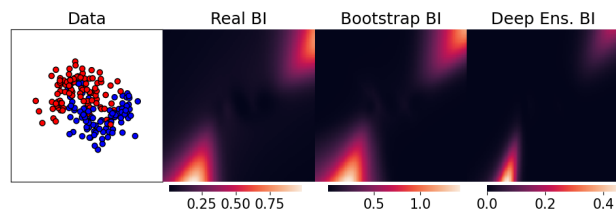

Figure 6: Comparing Bregman Information estimated via bootstrapping and Deep Ensembles for a neural network and a single training set with the ground truth ('Real BI').

a narrow decision boundary of high uncertainty even where no data is present. In contrast, the neural network and the naive Bayes classifier show increasing uncertainty towards out-of-domain regions along the decision boundary.

Figure 6 illustrates that Deep Ensembles and Bootstrapping (Efron, 1994) can serve as practical approximations for the ground truth Bregman Information. More simulations of toy tasks and MC Dropout (Gal & Ghahramani, 2016) for approximation are presented in Appendix C.

Based on our observations, the Bregman Information approximated by an ensemble could be a good proxy of the neural network's uncertainty even in regions we have not seen in the training data. We could differ between in-domain and out-of-domain instances, especially if the decision boundary in a high-dimensional space, like images, is 'open' towards multiple directions.

4.2 Out-of-Distribution Detection via Bregman Information

In this section, we use the Bregman Information of ensemble predictions for out-of-domain detection on image data. We propose a procedure along the following steps (c.f. Algorithm 1). First, we require an ensemble of predictions

to approximate the unknown uncertainty. Next, we compute the Bregman Information for each data instance in the validation set according to the ensemble predictions. We now have a set of instance-level Bregman Information for the in-domain data. Then, we compute the empirical quantiles of the Bregman Information values in the validation set. The central assumption is that out-of-domain data results in generally high Bregman Information. Consequently, thresholding our classification based on a chosen quantile should discard out-of-domain instances while only discarding a controlled amount of in-domain ones. For example, if we pick the 0.9-quantile, then 90% of the validation data is considered in-domain. In the ideal case, we only classify out-of-domain data instances close to the in-domain ones with a similar accuracy.

We assess the proposed procedure on CIFAR-10, ImageNet, and their corrupted versions (CIFAR-10-C and ImageNet-C) from Hendrycks & Dietterich (2019). These corruptions are provided in different types and levels of severity, ranging from one to five, where five is the worst corrupted (c.f. Appendix C). We evaluate Deep Ensembles and TTA as ensemble approaches. For ImageNet, we used pre-trained ResNet-50 models from Ashukha et al. (2020). To not skew the results through a performance gap between single models and model ensembles, we only use a single ResNet as classifier and use the ensemble only for estimating the Bregman Information.

As a baseline, we use Algorithm 1 with the predicted confidence score for uncertainty estimation (c.f. Algorithm 2 in Appendix C). The results are depicted in Figures 1 and 7. As we can see in the ImageNet case, upwards from the 0.6-quantile, our procedure successfully detects out-of-domain instances, which would have significantly lower accuracy than in-domain data. Consequently, the classified out-of-domain instances do not degrade the model performance. Further, the Bregman Information is very robust to the severity and the type of corruption. Contrary, the confidence score gives increasingly worse performance estimates with
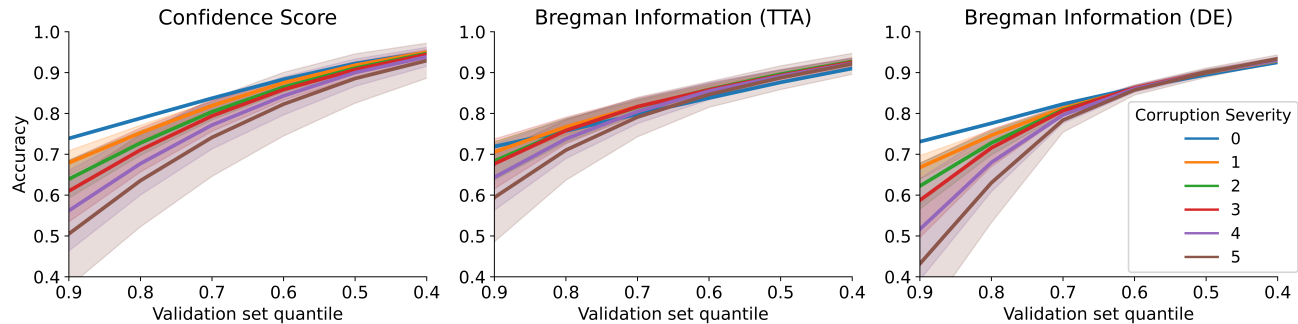
Figure 7: Accuracy after discarding for different types of uncertainty on ImageNet and for varying levels of corruption on ImageNet-C. The error bounds stem from the corruption type. The Bregman Information is more robust to corruption severity and corruption type for stricter thresholds compared to Confidence scores. **Left:** Using predicted confidence as uncertainty estimate. **Middle:** Using Bregman Information based on test-time augmentation for uncertainty. **Right:** Again, using Bregman Information but based on Deep Ensembles.

rising severity.

Algorithm 1 Classifying with uncertainty threshold

Require: Validation set $\mathscr{D}$, model, ensemble, $q \in [0, 1]$, test instance x'
  BIs ← [BI(ensemble(x)) for $x \in \mathscr{D}$]
  threshold ← quantile(BIs, q)
  if BI(ensemble(x')) > threshold **then**
    label as OOD   ▷ Warning in real-world application
  else
    return model(x')
  end if

4.3 Practical Limitations and Future Work

The main contribution in this work is of theoretical nature. Even though we demonstrate promising empirical results, more research is required for practical applications. We performed the experiments to suggest in what research areas the Bregman Information can potentially improve current methodologies, especially in the OOD setting. We leave a comparison in the context of state-of-the-art methods for future research. The different approximations of the Bregman Information are also preliminary. More extensive and thorough benchmarks may show better alternative approaches. Further, the used estimator $\hat{\mathbb{B}}_{\mathrm{LSE}}^{(n)}$ is only asymptotically unbiased, and underestimates the theoretical quantity. Corollary 3.6 and Section 3.5 may suggest to average ensembles in the logit space. But, as Gupta et al. (2022) demonstrated, averaging in the probability space may impact the bias in a positive way and reduce the overall error further than averaging in the logit space.

5 CONCLUSION

Through properties of functional Bregman divergences, we introduced a general bias-variance decomposition for proper scores. We discovered that the Bregman Information always represents the variance term. Our decomposition generalizes and provides new formulations of the exponential family and classification log-likelihood decomposition. Specifically, we formulated the classification case for logit predictions without requiring normalization to the log space. Further, we derived new general insights for ensemble predictions and how we construct confidence regions for predictions. As a real-world application, we proposed Bregman Information as uncertainty measure to facilitate model performance under domain drift via out-of-distribution detection for all degrees of severity and types of corruption.

References

Ashukha, A., Lyzhov, A., Molchanov, D., and Vetrov, D. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*, 2020.

Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., and Lafferty, J. Clustering with bregman divergences. *Journal of machine learning research*, 6(10), 2005.

Bregman, L. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200 − 217, 1967. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(67) 90040-7. URL http://www.sciencedirect.com/science/article/pii/0041555367900407.

Breiman, L. Random forests. *Machine learning*, 45(1):5–32, 2001.

Brinda, W., Klusowski, J. M., and Yang, D. Hölder's identity. *Statistics & Probability Letters*, 148:150–154, 2019.

Brofos, J., Shu, R., and Lederman, R. R. A bias-variance decomposition for bayesian deep learning. In *NeurIPS 2019 Workshop on Bayesian Deep Learning*, 2019.

Capiński, M. and Kopp, P. E. *Measure, integral and probability*, volume 14. Springer, 2004.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Csiszar, I. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *The annals of statistics*, 19(4):2032–2066, 1991.

Dawid, A. P. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, 2007.

Della Pietra, S., Della Pietra, V., and Lafferty, J. Duality and auxiliary functions for bregman distances (revised). Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA SCHOOL OF COMPUTER SCIENCE, 2002.

Domingos, P. A unified bias-variance decomposition. In *Proceedings of 17th international conference on machine learning*, pp. 231–238. Morgan Kaufmann Stanford, 2000.

Efron, B. *An introduction to the bootstrap*. CRC press, 1994.

Frigyik, B. A., Srivastava, S., and Gupta, M. R. Functional bregman divergence and bayesian estimation of distributions, 2006. URL https://arxiv.org/abs/cs/0611123.

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

Glorot, X., Bordes, A., and Bengio, Y. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323. JMLR Workshop and Conference Proceedings, 2011.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Grünwald, P. D. and Dawid, A. P. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *the Annals of Statistics*, 32(4):1367–1433, 2004.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Gupta, N., Smith, J., Adlam, B., and Mariet, Z. E. Ensembles of classifiers: a bias-variance perspective. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=lIOQFVncY9.

Haggenmüller, S., Maron, R. C., Hekler, A., Utikal, J. S., Barata, C., Barnhill, R. L., Beltraminelli, H., Berking, C., Betz-Stablein, B., Blum, A., Braun, S. A., Carr, R., Combalia, M., Fernandez-Figueras, M.-T., Ferrara, G., Fraitag, S., French, L. E., Gellrich, F. F., Ghoreschi, K., Goebeler, M., Guitera, P., Haenssle, H. A., Haferkamp, S., Heinzerling, L., Heppt, M. V., Hilke, F. J., Hobelsberger, S., Krahl, D., Kutzner, H., Lallas, A., Liopyris, K., Llamas-Velasco, M., Malvehy, J., Meier, F., Müller, C. S., Navarini, A. A., Navarrete-Dechent, C., Perasole, A., Poch, G., Podlipnik, S., Requena, L., Rotemberg, V. M., Saggini, A., Sangueza, O. P., Santonja, C., Schadendorf, D., Schilling, B., Schlaak, M., Schlager, J. G., Sergon, M., Sondermann,

W., Soyer, H. P., Starz, H., Stolz, W., Vale, E., Weyers, W., Zink, A., Krieghoff-Henning, E., Kather, J. N., von Kalle, C., Lipka, D. B., Fröhling, S., Hauschild, A., Kittler, H., and Brinker, T. J. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *European Journal of Cancer*, 156:202–216, 2021. ISSN 0959-8049. doi: https://doi.org/10.1016/j.ejca.2021.06.049. URL https://www.sciencedirect.com/science/article/pii/S0959804921004445.

Hansen, J. V. and Heskes, T. General bias/variance decomposition with target independent variance of error functions derived from the exponential family of distributions. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 2, pp. 207–210. IEEE, 2000.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hendrickson, A. D. and Buehler, R. J. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42(6):1916–1921, 1971.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Heskes, T. Bias/variance decompositions for likelihood-based estimators. *Neural Computation*, 10(6):1425–1433, 1998.

James, G. and Hastie, T. Generalizations of the bias/variance decomposition for prediction error. *Dept. Statistics, Stanford Univ., Stanford, CA, Tech. Rep*, 1997.

James, G. M. Variance and bias for general loss functions. *Machine learning*, 51(2):115–135, 2003.

Jones, L. K. and Byrne, C. L. General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis. *IEEE transactions on Information Theory*, 36(1):23–30, 1990.

Katsaouni, N., Tashkandi, A., Wiese, L., and Schulz, M. H. Machine learning based disease prediction from genotype data. *Biological Chemistry*, 402(8):871–885, 2021. doi: doi:10.1515/hsz-2021-0109. URL https://doi.org/10.1515/hsz-2021-0109.

Krizhevsky, A. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

Kurdila, A. J. and Zabarankin, M. *Convex functional analysis*. Springer Science & Business Media, 2006.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Murphy, K. P. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL probml.ai.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Ovcharov, E. Y. Existence and uniqueness of proper scoring rules. *J. Mach. Learn. Res.*, 16:2207–2230, 2015.

Ovcharov, E. Y. Proper scoring rules and Bregman divergence. *Bernoulli*, 24(1):53 – 79, 2018. doi: 10.3150/16-BEJ857. URL https://doi.org/10.3150/16-BEJ857.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pfau, D. A generalized bias-variance decomposition for bregman divergences. *Unpublished Manuscript*, 2013.

Rockafellar, R. T. *Convex analysis*, volume 18. Princeton university press, 1970.

Si, S., Tao, D., and Geng, B. Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):929–942, 2009.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=PxTIG12RRHS`.

Telgarsky, M. and Dasgupta, S. Agglomerative bregman clustering. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1011–1018, 2012.

Tomani, C. and Buettner, F. Towards trustworthy predictions from deep neural networks with fast adversarial calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9886–9896, 2021.

Ueda, N. and Nakano, R. Generalization error of ensemble estimators. In *Proceedings of International Conference on Neural Networks (ICNN'96)*, volume 1, pp. 90–95. IEEE, 1996.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Yang, Z., Yu, Y., You, C., Steinhardt, J., and Ma, Y. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pp. 10767–10777. PMLR, 2020.

Yen, M.-H., Liu, D.-W., Hsin, Y.-C., Lin, C.-E., and Chen, C.-C. Application of the deep learning for the prediction of rainfall in southern taiwan. *Scientific reports*, 9(1):1–9, 2019.

A OVERVIEW

In this appendix, we provide more rigorous definitions than in the main paper and the missing proofs in Appendix B. Further, we give a more detailed description of the experiments and showcase additional empirical results in Appendix C.

B MISSING PROOFS

We first provide some more rigorous definitions than in the main paper in Section B.1. There, we also introduce and prove some basic facts with respect to these definitions, which we will then use as tools in the proofs of Lemma 3.1 in Section B.2, Theorem 3.2 in Section B.3, Proposition 3.4 in Section B.4, Corollary 3.6 in Section B.5, and Proposition 3.8 in Section B.6.

B.1 Preliminaries

The following definitions are based on (Ovcharov, 2018). Let $\mathscr{P}$ be a convex set of probability measures of a measure space $(\Omega, \mathscr{F}, \mu)$. We define the space of finite linear combinations of $\mathscr{P}$ as $\mathrm{span}\mathscr{P} = \left\{ \sum_{i=1}^{n} a_i P_i \mid a_1, \ldots, a_n \in \mathbb{R}, P_1, \ldots, P_n \in \mathscr{P}, n \in \mathbb{N} \right\}$. We define the space of $\mathscr{P}$-integrable functions as $\mathscr{L}(\mathscr{P}) = \left\{ f \mid \int |f| \, \mathrm{d}P < \infty, P \in \mathscr{P} \right\}$. We use $\cdot \cdot \cdot : \mathscr{L}(\mathscr{P}) \times \mathrm{span}\mathscr{P} \to \mathbb{R}$ defined as $f \cdot P = \int f \mathrm{d}P$ as the pairing between the dual spaces $\mathrm{span}\mathscr{P}$ and $\mathscr{L}(\mathscr{P})$.

For completeness, we repeat the following definition from the main paper in a more extensive form.

Definition B.1. Let $\phi: U \to \mathbb{R}$ be a convex function. The **subdifferential** of ϕ at a point $P \in U$ is defined as $\partial\phi(P) = \left\{ P^* \in \mathscr{L}(\mathscr{P}) \mid \phi(Q) \geq \phi(P) + P^* \cdot (Q - P), Q \in U \right\}$. An element $P^* \in \partial\phi(P)$ is called **subgradient** of ϕ at P. We call a function $\phi': U \to \mathscr{L}(\mathscr{P})$ defined as $\phi'(P) = P^* \in \partial\phi(P)$ a **subgradient** of ϕ on U.

Note that the inequality for the subgradient becomes strict if ϕ is strictly convex and $P \neq Q$.

Since the definition of the convex conjugate in the main paper is rather informal, we state a more rigorous version (Rockafellar, 1970).

Definition B.2. Given a vector space X with dual vector space X^*, pairing $\langle x, x^* \rangle = x(x^*)$ for $x \in X$ and $x^* \in X^*$, and a function $\phi: X \to \mathbb{R} \cup \{-\infty, \infty\}$, the convex conjugate $\phi^*: X^* \to \mathbb{R} \cup \{-\infty, \infty\}$ of ϕ is defined as $\phi^*(x^*) = \sup_{x \in X} \langle x, x^* \rangle - \phi(x)$.

In our context, we have $X = \mathrm{span}\mathscr{P}$, $X^* = \mathscr{L}(\mathscr{P})$, and $\langle x, x^* \rangle = x^* \cdot x$. To fit the definition, note that we may always extend a convex function $\phi: \mathrm{span}\mathscr{P} \supset U \to \mathbb{R}$ to $\mathrm{span}\mathscr{P}$ via $\phi(x) = \infty$ for $x \notin U$. We imply this throughout this work. We now state the following, which will be used in later proofs.

Lemma B.3. *Let P^* be a subgradient at point P of a convex function $\phi: U \to \mathbb{R}$. Then it holds that*

$$\phi^*(P^*) = P^* \cdot P - \phi(P). \tag{7}$$

Proof. By definition we have for $P, Q \in U$ and subgradient P^* of $\phi(P)$ that

$$\phi(Q) \geq \phi(P) + P^* \cdot (Q - P), \tag{8}$$

from which follows

$$P^* \cdot P - \phi(P) \geq P^* \cdot Q - \phi(Q). \tag{9}$$

Since $P \in \mathrm{span}\mathscr{P}$, we have

$$\phi^*(P^*) = \sup_{Q \in \mathrm{span}\mathscr{P}} P^* \cdot Q - \phi(Q) = P^* \cdot P - \phi(P). \tag{10}$$

$\square$

Further, we will make use of the following properties.

Lemma B.4. *If $\phi: U \to \mathbb{R}$ is strictly convex, then any subgradient ϕ' of ϕ is injective and its inverse $(\phi')^{-1}$ exists on $\phi'(U) := \{\phi'(P) \mid P \in U\}$. Further, $(\phi')^{-1}$ is a subgradient of the convex conjugate ϕ^* on $\phi'(U)$.*

Proof. The proof is similar to the proof of Theorem 6.2.1 from Kurdila & Zabarankin (2006).

For $P, Q \in U$ with $P \neq Q$ we have

$$
\begin{aligned}
&\phi(Q) > \phi(P) + \phi'(P) \cdot (Q - P) \\
\Longrightarrow\ &\phi(Q) - \phi(P) > \phi'(P) \cdot (Q - P)
\end{aligned}
\tag{11}
$$

Reversing the roles of P and Q also gives

$$
\phi(P) - \phi(Q) > -\phi'(Q) \cdot (Q - P). \tag{12}
$$

Adding the LHS and RHS of the last two inequalities results in

$$
\begin{aligned}
&0 > \left(\phi'(P) - \phi'(Q)\right) \cdot (Q - P) \\
\Longrightarrow\ &\phi'(P) \neq \phi'(Q).
\end{aligned}
\tag{13}
$$

Consequently, since $\phi'(P)$ is unique for each $P \in U$, it is injective, and the inverse $\left(\phi'\right)^{-1}$ exists for all $P^* \in \phi'(U)$.

Next, we show that $\left(\phi'\right)^{-1}$ is a subgradient of ϕ^* on $\phi'(U)$. By definition, for all $P^* \in \phi'(U)$ there exists $P \in U$ such that $P^* = \phi'(P)$ and $\left(\phi'\right)^{-1}\left(P^*\right) = P$. For all $P^*, Q^* \in \phi'(U)$ we have

$$
\begin{aligned}
&\phi^*\left(Q^*\right) \geq \phi^*\left(P^*\right) + \left(Q^* - P^*\right) \cdot \left(\phi'\right)^{-1}\left(P^*\right) = \phi^*\left(P^*\right) + \left(Q^* - P^*\right) \cdot P \\
\Longleftrightarrow\ &\sup_{Y \in \operatorname{span}\mathscr{P}} Q^* \cdot Y - \phi(Y) \geq \sup_{X \in \operatorname{span}\mathscr{P}} P^* \cdot X - \phi(X) + Q^* \cdot P - P^* \cdot P \\
\overset{\text{Le B.3}}{\Longleftrightarrow}\ &Q^* \cdot Q - \phi(Q) \geq P^* \cdot P - \phi(P) + Q^* \cdot P - P^* \cdot P \\
\Longleftrightarrow\ &\phi(P) + Q^* \cdot Q \geq \phi(Q) + Q^* \cdot P \\
\Longleftrightarrow\ &\phi(P) \geq \phi(Q) + Q^* \cdot (P - Q).
\end{aligned}
\tag{14}
$$

Since $Q^* = \phi'(Q)$ is a subgradient of ϕ at point Q, the last line holds and confirms that $\left(\phi'\right)^{-1}$ is a subgradient of ϕ^* on $\phi'(U)$.

$\square$

So far, we established the theoretical foundation to perform the flipping in the argument of a functional Bregman divergence. But, Lemma 3.1 also requires that the convex conjugate is finite on $\mathbb{E}\left[Q^*\right]$. This is not self evident since $\mathbb{E}\left[Q^*\right] \notin \phi'(U)$ in general. Consequently, we also require the following.

Lemma B.5. *If Q^* is a random variable with values in $\mathscr{L}(\mathscr{P})$, such that $\mathbb{E}\left[\phi^*\left(Q^*\right)\right]$ is finite and $\mathbb{E}\left[Q^*\right] \in \mathscr{L}(\mathscr{P})$, then $\phi^*\left(\mathbb{E}\left[Q^*\right]\right) \in \mathbb{R}$ for any convex function $\phi: U \to \mathbb{R}$.*

Proof. By using Jensen's inequality we get

$$
\infty > \mathbb{E}\left[\phi^*\left(Q^*\right)\right] \geq \phi^*\left(\mathbb{E}\left[Q^*\right]\right). \tag{15}
$$

Further, due to $\mathbb{E}\left[Q^*\right] \in \mathscr{L}(\mathscr{P})$, for any $P \in U$ we have

$$
-\infty < \mathbb{E}\left[Q^*\right] \cdot P - \phi(P) \leq \sup_{Q \in \operatorname{span}\mathscr{P}} \mathbb{E}\left[Q^*\right] \cdot Q - \phi(Q) = \phi^*\left(\mathbb{E}\left[Q^*\right]\right). \tag{16}
$$

$\square$

We can now state Lemma 3.1 and its proof in a more rigorous way than in the main paper.

B.2   Proof of Lemma 3.1

Let ϕ' be the subgradient of a strictly convex function $\phi\colon U \to \mathbb{R}$. We first show the first equality in Lemma 3.1. Based on the definition of a functional Bregman divergence with $p, q \in U$, we have

$$
\begin{aligned}
&d_{\phi,\phi'}\left(p, q\right) \\
&= \phi\left(q\right) - \phi\left(p\right) - \phi'\left(p\right) \cdot \left(q - p\right) \\
&= \phi\left(q\right) - \phi\left(p\right) - \phi'\left(p\right) \cdot q + \phi'\left(p\right) \cdot p + \phi'\left(q\right) \cdot q - \phi'\left(q\right) \cdot q \\
&\overset{\text{Le } B.3}{=} \phi^*\left(\phi'\left(p\right)\right) - \phi^*\left(\phi'\left(q\right)\right) - \left(\phi'\left(p\right) - \phi'\left(q\right)\right) \cdot q \\
&\overset{\text{Le } B.4}{=} \phi^*\left(\phi'\left(p\right)\right) - \phi^*\left(\phi'\left(q\right)\right) - \left(\phi'\left(p\right) - \phi'\left(q\right)\right) \cdot \left(\phi'\right)^{-1}\left(\phi'\left(q\right)\right) \\
&= d_{\phi^*,(\phi')^{-1}}\left(\phi'\left(q\right), \phi'\left(p\right)\right).
\end{aligned}
\tag{17}
$$

In the last line, we used the definition of functional Bregman divergences and the fact that $\left(\phi'\right)^{-1}$ is a subgradient of ϕ^* (c.f. Lemma B.4). If $p^*, q^* \in \phi'\left(U\right)$, we can set $\left(\phi'\right)^{-1}\left(p^*\right) = p$ and $\left(\phi'\right)^{-1}\left(q^*\right) = q$ in Equation 17 and receive a similar result for the second equality in Lemma 3.1.

So far, we showed that $d_{\phi^*,(\phi')^{-1}}\colon \phi'\left(U\right) \times \phi'\left(U\right) \to \mathbb{R}$ is well-defined, but in general $\mathbb{E}\left[Q^*\right] \notin \phi'\left(U\right)$. Let $\partial\phi\left(U\right) := \bigcup_{P \in U} \partial\phi\left(P\right)$ be the set of all subgradients of ϕ and Conv the convex hull operator. From Lemma B.5 also follows that for all $C \in \mathrm{Conv}\left(\partial\phi\left(U\right)\right)$ the convex conjugate $\phi^*\left(C\right)$ is finite. This can be seen by restricting Q^* to binary random variables mapping into $\partial\phi\left(U\right)$. Consequently, the divergence $d_{\phi^*,(\phi')^{-1}}\colon \phi'\left(U\right) \times \mathrm{Conv}\left(\partial\phi\left(U\right)\right) \to \mathbb{R}$ is well-defined and finite. We implied this domain in the last equation in Lemma 3.1. A divergences with different domains for its arguments is in a strictly sense not a (functional) Bregman divergence, but we can always extend the first domain by extending $\left(\phi'\right)^{-1}$. We denote this non-unique extension with $\phi^{*\prime}\colon \mathrm{Conv}\left(\partial\phi\left(U\right)\right) \to U$. Then, $d_{\phi^*,\phi^{*\prime}}\colon \mathrm{Conv}\left(\partial\phi\left(U\right)\right) \times \mathrm{Conv}\left(\partial\phi\left(U\right)\right) \to \mathbb{R}$ is a functional Bregman divergence. Note also that the functional Bregman Information is invariant towards the subgradient:

$$
\mathbb{B}_{\phi^*}\left[Q^*\right] = \mathbb{E}\left[d_{\phi^*,\phi^{*\prime}}\left(\mathbb{E}\left[Q^*\right], Q^*\right)\right] = \mathbb{E}\left[\phi^*\left(Q^*\right)\right] - \phi^*\left(\mathbb{E}\left[Q^*\right]\right).
\tag{18}
$$

Thus, the domain of the first argument is just a technicality in our context and can be ignored in practical applications in favor of a unique $\left(\phi'\right)^{-1}$.

Now, let Q^* be a random variable with realizations in $\phi'\left(U\right)$ such that $\mathbb{E}\left[\phi^*\left(Q^*\right)\right]$ exists and $\mathbb{E}\left[Q^*\right] \in \mathscr{L}\left(\mathscr{P}\right)$. Then, we get the last equality in Lemma 3.1 with $p^* \in \phi'\left(U\right)$ by

$$
\begin{aligned}
&\mathbb{E}\left[d_{\phi^*,(\phi')^{-1}}\left(p^*, Q^*\right)\right] \\
&= \mathbb{E}\left[\phi^*\left(Q^*\right) - \phi^*\left(p^*\right) - \left(Q^* - p^*\right) \cdot \left(\phi'\right)^{-1}\left(p^*\right)\right] \\
&\overset{\text{Le } B.5}{=} \mathbb{E}\left[\phi^*\left(Q^*\right) - \phi^*\left(p^*\right) - \left(Q^* - p^*\right) \cdot \left(\phi'\right)^{-1}\left(p^*\right)\right] + \phi^*\left(\mathbb{E}\left[Q^*\right]\right) - \phi^*\left(\mathbb{E}\left[Q^*\right]\right) \\
&= \phi^*\left(\mathbb{E}\left[Q^*\right]\right) - \phi^*\left(p^*\right) - \left(\mathbb{E}\left[Q^*\right] - p^*\right) \cdot \left(\phi'\right)^{-1}\left(p^*\right) + \mathbb{E}\left[\phi^*\left(Q^*\right)\right] - \phi^*\left(\mathbb{E}\left[Q^*\right]\right) \\
&= d_{\phi^*,(\phi')^{-1}}\left(p^*, \mathbb{E}\left[Q^*\right]\right) + \mathbb{B}_{\phi^*}\left[Q^*\right].
\end{aligned}
\tag{19}
$$

We now have the necessary tools to prove our main result.

B.3   Proof of Theorem 3.2

For completeness, we derive the relation between proper scores and functional Bregman divergences in the following, even though it is already known in the literature (Ovcharov, 2018).

Note that a score S proper on $U \subset \mathscr{P}$ is a subgradient of G on U since for all $P, Q \in U$

$$
\begin{aligned}
&S\left(Q\right) \cdot Q \geq S\left(P\right) \cdot Q \\
&\Longleftrightarrow S\left(Q\right) \cdot Q \geq S\left(P\right) \cdot P + S\left(P\right) \cdot Q - S\left(P\right) \cdot P \\
&\overset{\text{def}}{\Longleftrightarrow} G\left(Q\right) \geq G\left(P\right) + S\left(P\right) \cdot \left(Q - P\right).
\end{aligned}
\tag{20}
$$

The relation between a proper score S and a functional Bregman divergence $d_{G,S}$ on U is then given by

$$
\begin{aligned}
S(P) \cdot Q &= S(Q) \cdot Q - S(Q) \cdot Q + S(P) \cdot Q - S(P) \cdot P + S(P) \cdot P \\
&\stackrel{\text{def}}{=} G(Q) - G(Q) + G(P) + S(P) \cdot (Q - P) \\
&\stackrel{\text{def}}{=} G(Q) - d_{G,S}(P, Q).
\end{aligned}
\tag{21}
$$

Now, let S be strictly proper on U. Then its negative entropy G is strictly convex on U (Ovcharov, 2018). Further, if for all $Q \in U$ and a random variable $P \colon \Omega \to U$ the integrals $\mathbb{E}[S(P)(Y)]$, $\mathbb{E}[d_{G,S}(P,Q)]$, and $\mathbb{E}[G^*(S(P))]$ exist, as well as $\mathbb{E}[S(P)] \in \mathscr{L}(\mathscr{P})$, then we have

$$
\begin{aligned}
\mathbb{E}[-S(P)(Y)] &= -\mathbb{E}[S(P) \cdot Q] \\
&\stackrel{\text{Eq (21)}}{=} -G(Q) + \mathbb{E}[d_{G,S}(P,Q)] \\
&\stackrel{\text{Le 3.1}}{=} -G(Q) + \mathbb{E}[d_{G^*,S^{-1}}(S(Q), S(P))] \\
&\stackrel{\text{Le 3.1}}{=} -G(Q) + d_{G^*,S^{-1}}(S(Q), \mathbb{E}[S(P)]) + \mathbb{B}_{G^*}[S(P)].
\end{aligned}
\tag{22}
$$

B.4 Proof of Proposition 3.4

Theorem 3.2 is stated for distributions. Exponential families are usually stated in form of their density or mass function, which are also used for the log-likelihood. Further, Proposition 3.4 assumes we are restricted to a specific exponential family. In this context, the Radon-Nikodym derivative of a distribution P_θ is $p_\theta := \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}$ with base measure μ of the related measure space $(\Omega, \mathscr{F}, \mu)$. For continuous distributions, μ is the Lebesgue measure, and for discrete distributions, μ is the counting measure. We assume the set of distributions $\mathscr{P}$ consists of distributions with the same base measure. To state our proof for discrete as well as continuous families, we will use the Radon-Nikodym formulation.

Further, we require the more general formulations for the log score, the negative Shannon entropy, and the log partition function by $S(P) = \ln \frac{\mathrm{d}P}{\mathrm{d}\mu}$, $H(P) = \ln \frac{\mathrm{d}P}{\mathrm{d}\mu} \cdot P = \int_\Omega \ln \frac{\mathrm{d}P}{\mathrm{d}\mu} \mathrm{d}P$, and $H^*(P^*) = \ln(\exp P^* \cdot \mu) = \ln \int_\Omega \exp P^* \mathrm{d}\mu$. For densities, these formulations reduce to the ones provided in Example 3.3.

For an exponential family, we have

$$
\frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}(x) = p_\theta(x) = \exp(\langle \theta, T(x) \rangle - A(\theta)) h(x)
\tag{23}
$$

and, thus, also

$$
\frac{\mathrm{d}P_\theta}{\mathrm{d}\mu} = p_\theta = \exp(\langle \theta, T \rangle - A(\theta)) h \in \mathscr{L}(P).
\tag{24}
$$

The last statement also introduces the notation we will use.

For the log score, it follows

$$
\mathbb{E}[S(P_{\hat\theta})] = \mathbb{E}\left[\ln \frac{\mathrm{d}P_{\hat\theta}}{\mathrm{d}\mu}\right] = \langle \mathbb{E}[\hat\theta], T \rangle - \mathbb{E}[A(\hat\theta)] - \ln h
\tag{25}
$$

which gives

$$
\begin{aligned}
H^*\left(\mathbb{E}\left[\ln \frac{\mathrm{d}P_{\hat\theta}}{\mathrm{d}\mu}\right]\right) &= \ln \int \exp\left(\langle \mathbb{E}[\hat\theta], T \rangle - \mathbb{E}[A(\hat\theta)]\right) h \mathrm{d}\mu \\
&= \ln \int \exp\left(\langle \mathbb{E}[\hat\theta], T \rangle\right) h \mathrm{d}\mu - \mathbb{E}[A(\hat\theta)] \\
&= A(\mathbb{E}[\hat\theta]) - \mathbb{E}[A(\hat\theta)].
\end{aligned}
\tag{26}
$$

Consequently, with $\mathbb{B}_{H^*}\left[\ln\frac{\mathrm{d}P_{\hat{\theta}}}{\mathrm{d}\mu}\right] = -H^*\left(\mathbb{E}\left[\ln\frac{\mathrm{d}P_{\hat{\theta}}}{\mathrm{d}\mu}\right]\right)$ stated in Example 3.3 we can already say

$$\mathbb{B}_{H^*}\left[\ln\frac{\mathrm{d}P_{\hat{\theta}}}{\mathrm{d}\mu}\right] = \mathbb{E}\left[A\left(\hat{\theta}\right)\right] - A\left(\mathbb{E}\left[\hat{\theta}\right]\right) = \mathbb{B}_A\left[\hat{\theta}\right]. \tag{27}$$

The reduction from a functional Bregman Information to a vector-based Bregman Information is a remarkable fact for exponential families, which will not hold for the bias term as we will see in the following. For the bias term, we first have to make some further additional statements. Note that from definition of the log score, it follows that $S^{-1}\left(P^*\right) = \int\exp P^*\mathrm{d}\mu$, which is a mapping from $\mathscr{F}$ to $\mathbb{R}$. To confirm the inverse, note that for all $P \in \mathscr{P}$ and for all $F \in \mathscr{F}$, we have

$$S^{-1}\left(S\left(P\right)\right)\left(F\right) = \left(\int\exp\ln\frac{\mathrm{d}P}{\mathrm{d}\mu}\mathrm{d}\mu\right)\left(F\right) = \int_F\frac{\mathrm{d}P}{\mathrm{d}\mu}\mathrm{d}\mu \overset{\text{(i)}}{=} P\left(F\right), \tag{28}$$

where we used the Radon-Nikodym theorem in (i).

Then, for all $P^* \in \{S\left(P\right) \mid P \in \mathscr{P}\} \subset \mathscr{L}\left(\mathscr{P}\right)$ and $x \in \Omega$ it holds almost surely

$$\begin{aligned}
S\left(S^{-1}\left(P^*\right)\right)\left(x\right) &= \left(\ln\frac{\mathrm{d}\int\exp P^*\mathrm{d}\mu}{\mathrm{d}\mu}\right)\left(x\right) \\
&= \ln\left(\frac{\mathrm{d}\int\exp P^*\mathrm{d}\mu}{\mathrm{d}\mu}\left(x\right)\right) \\
&= \ln\left(\lim_{B\to\{x\}}\frac{1}{\mu\left(B\right)}\int_B\exp P^*\mathrm{d}\mu\right) \\
&\overset{\text{(ii)}}{=} \ln\left(\exp P^*\left(x\right)\right) = P^*\left(x\right),
\end{aligned} \tag{29}$$

where we used the Lebesgue differentiation theorem in (ii).

Remark B.6. It is possible to extend S^{-1} via $\frac{1}{\int_\Omega\exp P^*\mathrm{d}\mu}S^{-1}$ to $\mathscr{L}\left(\mathscr{P}\right)$. This makes it the subgradient of H^* on $\mathscr{L}\left(\mathscr{P}\right)$ but it is unnecessary for the proof.

Following from Equation (25), we will also make use of

$$\int_\Omega\mathbb{E}\left[\ln\frac{\mathrm{d}P_{\hat{\theta}}}{\mathrm{d}\mu}\right]\mathrm{d}Q = \int_\Omega\left\langle\mathbb{E}\left[\hat{\theta}\right], T\right\rangle - \mathbb{E}\left[A\left(\hat{\theta}\right)\right] - \ln h \;\mathrm{d}Q \overset{Y\sim Q}{=} \left\langle\mathbb{E}\left[\hat{\theta}\right], \mathbb{E}\left[T\left(Y\right)\right]\right\rangle - \mathbb{E}\left[A\left(\hat{\theta}\right)\right] - \mathbb{E}\left[\ln h\left(Y\right)\right]. \tag{30}$$

For the bias term in Theorem 3.2, we first assume a general $Y \sim Q$ to demonstrate our claim about Proposition 3.4 that the decomposition holds even when the distribution assumption is wrong. Now, we can state that

$$d_{H^*,S^{-1}}\left(\ln\frac{dQ}{d\mu}, \mathbb{E}\left[\ln\frac{dP_{\hat{\theta}}}{d\mu}\right]\right)$$

$$\overset{\text{def}}{=} \underbrace{H^*\left(\mathbb{E}\left[\ln\frac{dP_{\hat{\theta}}}{d\mu}\right]\right)}_{\overset{\text{Eq (26)}}{=} A(\mathbb{E}[\hat{\theta}]) - \mathbb{E}[A(\hat{\theta})]} - \underbrace{H^*\left(\ln\frac{dQ}{d\mu}\right)}_{\overset{\text{Ex } 3.3}{=} 0} - \left(\mathbb{E}\left[\ln\frac{dP_{\hat{\theta}}}{d\mu}\right] - \ln\frac{dQ}{d\mu}\right)\cdot\underbrace{S^{-1}\left(\ln\frac{dQ}{d\mu}\right)}_{\overset{\text{Eq (28)}}{=} Q}$$

$$= A\left(\mathbb{E}[\hat{\theta}]\right) - \mathbb{E}\left[A(\hat{\theta})\right] - \int_{\Omega}\mathbb{E}\left[\ln\frac{dP_{\hat{\theta}}}{d\mu}\right]dQ + H(Q)$$

$$\overset{\text{Eq (30)}}{=} A\left(\mathbb{E}[\hat{\theta}]\right) - \left\langle\mathbb{E}[\hat{\theta}], \mathbb{E}[T(Y)]\right\rangle + \mathbb{E}[\ln h(Y)] + H(Q)$$

$$= A\left(\mathbb{E}[\hat{\theta}]\right) - \left\langle\mathbb{E}[\hat{\theta}], \mathbb{E}[T(Y)]\right\rangle + \mathbb{E}[\ln h(Y)] + H(Q) + A^*(\mathbb{E}[T(Y)]) - A^*(\mathbb{E}[T(Y)])$$

$$\overset{\text{Le } B.3}{=} A\left(\mathbb{E}[\hat{\theta}]\right) - \left\langle\mathbb{E}[\hat{\theta}], \mathbb{E}[T(Y)]\right\rangle + \mathbb{E}[\ln h(Y)] + H(Q) +$$
$$\quad + \left\langle\nabla A(\mathbb{E}[T(Y)]), \mathbb{E}[T(Y)]\right\rangle - A(\nabla A^*(\mathbb{E}[T(Y)])) - A^*(\mathbb{E}[T(Y)])$$

$$= A\left(\mathbb{E}[\hat{\theta}]\right) - A(\nabla A^*(\mathbb{E}[T(Y)])) - \left\langle\nabla A(\nabla A^*(\mathbb{E}[T(Y)])), \mathbb{E}[\hat{\theta}] - A^*(\mathbb{E}[T(Y)])\right\rangle +$$
$$\quad + \mathbb{E}[\ln h(Y)] + H(Q) - A^*(\mathbb{E}[T(Y)])$$

$$\overset{\text{def}}{=} d_A\left(\nabla A^*(\mathbb{E}[T(Y)]), \mathbb{E}[\hat{\theta}]\right) + \mathbb{E}[\ln h(Y)] + H(Q) - A^*(\mathbb{E}[T(Y)]).$$

(31)

As we can see, while the functional Bregman Information nicely reduces to a vector-based Bregman Information, it is not the case for the functional form of the bias. Specifically, the functional bias and noise term have to be taken together to end up with a vector-based bias term.

So far, Y was arbitrarily distributed, but we require an additional restriction to end up with the formulation in Proposition 3.4. If we assume that $Y \sim Q = P_{\theta}$ follows a distribution from the respective exponential family with natural parameter θ, then we have $\nabla A^*(\mathbb{E}[T(Y)]) = \theta$, which gives in the last line in Equation (31) that

$$d_{H^*,S^{-1}}\left(\ln\frac{dQ}{d\mu}, \mathbb{E}\left[\ln\frac{dP_{\hat{\theta}}}{d\mu}\right]\right) = d_A\left(\theta, \mathbb{E}[\hat{\theta}]\right) + \mathbb{E}[\ln h(Y)] + H(Q) - A^*(\nabla A(\theta)).$$

(32)

B.5   Proof of Corollary 3.6

We now provide proof for the closed-form decomposition of the classification log-likelihood. Since it corresponds to the log-likelihood for the categorical distribution (an exponential family), we can directly derive it from Proposition 3.4.

For the categorical distribution with k classes, we have for $\theta \in \Theta = \mathbb{R}^{k-1}$ the log-partition $A(\theta) = \ln\left(1 + \sum_i^{k-1}\exp\theta_i\right)$ and $h \equiv 1$. The gradient is $\nabla A(\theta) = \frac{1}{1+\sum_i^{k-1}\exp\theta_i}(\exp\theta_1, \ldots, \exp\theta_{k-1})^{\top}$ Further, we have for $\theta^* \in \Theta^* = \left\{(p_1, \ldots, p_{k-1})^{\top} \mid p_1, \ldots, p_{k-1} \in (0,1), \sum_i p_i < 1\right\}$ the convex conjugate $A^*(\theta^*) = \left(1 - \sum_{i=1}^{k-1}\theta_i^*\right)\ln\left(1 - \sum_{i=1}^{k-1}\theta_i^*\right) + \sum_{i=1}^{k-1}\theta_i^*\ln\theta_i^*$ with $\nabla A^*(\theta^*) = \left(\ln\frac{\theta_1^*}{1-\sum_{i=1}^{k-1}\theta_i^*}, \ldots, \ln\frac{\theta_{k-1}^*}{1-\sum_{i=1}^{k-1}\theta_i^*}\right)^{\top}$.

Further, we will relate each $\theta \in \mathbb{R}^{k-1}$ to an equivalence class

$$[\theta] := \left\{z \in \mathbb{R}^k \mid z_1 = \theta_1 + z_k, \ldots, z_{k-1} = \theta_{k-1} + z_k\right\} = \left\{z \in \mathbb{R}^k \mid \mathrm{sm}\left((\theta_1, \ldots, \theta_{k-1}, 0)^{\top}\right) = \mathrm{sm}(z)\right\}.$$

(33)

All members of an equivalence class give the same softmax output. Now, for any $\theta, \hat{\theta} \in \Theta$ and $z \in [\theta], \hat{z} \in [\hat{\theta}]$, it holds that

$$
\begin{aligned}
\mathbb{B}_A\left[\hat{\theta}\right] &= \mathbb{E}\left[A\left(\hat{\theta}\right)\right] - A\left(\mathbb{E}\left[\hat{\theta}\right]\right) \\
&= \mathbb{E}\left[\ln\left(1 + \sum_{i=1}^{k-1} \exp\hat{\theta}_i\right)\right] - \ln\left(1 + \sum_{i=1}^{k-1} \exp\mathbb{E}\left[\hat{\theta}_i\right]\right) \\
&= \mathbb{E}\left[\ln\left(1 + \sum_{i=1}^{k-1} \exp\hat{\theta}_i\right)\right] - \ln\left(1 + \sum_{i=1}^{k-1} \exp\mathbb{E}\left[\hat{\theta}_i\right]\right) + \mathbb{E}\left[\ln\exp\hat{z}_k\right] - \ln\exp\mathbb{E}\left[\hat{z}_k\right] \\
&= \mathbb{E}\left[\ln\left(\exp\hat{z}_k + \sum_{i=1}^{k-1} \exp\left(\hat{\theta}_i + \hat{z}_k\right)\right)\right] - \ln\left(\exp\mathbb{E}\left[\hat{z}_k\right] + \sum_{i=1}^{k-1} \exp\mathbb{E}\left[\hat{\theta}_i + \hat{z}_k\right]\right) \\
&= \mathbb{E}\left[\ln\sum_{i=1}^{k} \exp\hat{z}_i\right] - \ln\sum_{i=1}^{k} \exp\mathbb{E}\left[\hat{z}_i\right] \\
&= \mathbb{B}_{\mathrm{LSE}}\left[\hat{z}\right].
\end{aligned}
\tag{34}
$$

For the bias term, it holds that

$$
\begin{aligned}
d_A\left(\theta, \mathbb{E}\left[\hat{\theta}\right]\right) &\overset{\text{def}}{=} \ln\left(1 + \sum_{i=1}^{k-1} \exp\mathbb{E}\left[\hat{\theta}_i\right]\right) - \ln\left(1 + \sum_{i=1}^{k-1} \exp\theta_i\right) - \sum_{i=1}^{k-1} \frac{\exp\theta_i}{1 + \sum_{j=1}^{k-1} \exp\theta_j}\left(\mathbb{E}\left[\hat{\theta}_i\right] - \theta_i\right) \\
&= \ln\left(1 + \sum_{i=1}^{k-1} \exp\mathbb{E}\left[\hat{z}_i - \hat{z}_k\right]\right) - \ln\sum_{i=1}^{k} \exp z_i - \sum_{i=1}^{k-1} \frac{\exp z_i}{\sum_{j=1}^{k} \exp z_j}\left(\mathbb{E}\left[\hat{z}_i - \hat{z}_k\right] - z_i\right) \\
&= \ln\sum_{i=1}^{k} \exp\mathbb{E}\left[\hat{z}_i\right] - \mathbb{E}\left[\hat{z}_k\right] - \ln\sum_{i=1}^{k} \exp z_i - \sum_{i=1}^{k} \mathrm{sm}_i\left(z\right)\left(\mathbb{E}\left[\hat{z}_i\right] - z_i\right) + \underbrace{\sum_{i=1}^{k} \mathrm{sm}_i\left(z\right)\mathbb{E}\left[\hat{z}_k\right]}_{=1} \\
&= \ln\sum_{i=1}^{k} \exp\mathbb{E}\left[\hat{z}_i\right] - \ln\sum_{i=1}^{k} \exp z_i - \sum_{i=1}^{k} \mathrm{sm}_i\left(z\right)\left(\mathbb{E}\left[\hat{z}_i\right] - z_i\right) \\
&= \mathrm{LSE}\left(\mathbb{E}\left[\hat{z}\right]\right) - \mathrm{LSE}\left(z\right) - \left\langle\nabla\mathrm{LSE}\left(z\right), \mathbb{E}\left[\hat{z}\right] - z\right\rangle \\
&\overset{\text{def}}{=} d_{\mathrm{LSE}}\left(z, \mathbb{E}\left[\hat{z}\right]\right).
\end{aligned}
\tag{35}
$$

For $i \in \{1, \ldots, k\}$, we use the probability mass function $Q_i := \frac{\mathrm{d}Q}{\mathrm{d}\mu}(i)$ of the distribution Q with counting measure μ for shorter notations. Last, the noise term gives

$$
-A^*\left(\nabla A\left(\theta\right)\right) = -A^*\left(\left(Q_1, \ldots, Q_{k-1}\right)^\intercal\right) = -\sum_{i=1}^{k-1} Q_i \ln Q_i - \left(1 - \sum_{i=1}^{k-1} Q_i\right)\ln\left(1 - \sum_{i=1}^{k-1} Q_i\right) = H\left(Q\right).
\tag{36}
$$

Let $\mathrm{sm}^{-1}\left(p\right) := \left(\ln\frac{p_1}{p_k}, \ldots, \ln\frac{p_{k-1}}{p_k}, 0\right)^\intercal$ for a probability vector p. Further, let Q have the natural parameter vector θ, which gives $Q = \mathrm{sm}\left(z\right)$ for $z \in [\theta]$ and $\mathrm{sm}^{-1}\left(Q\right) \in [\theta]$. Using the Equations (34), (35), and (36) with Corollary 3.6, we then receive for $Y \sim Q$ and $z \in [\theta], \hat{z} \in \left[\hat{\theta}\right]$

$$
\begin{aligned}
\mathbb{E}\left[-\ln\mathrm{sm}_Y\left(\hat{z}\right)\right] &= \mathbb{E}\left[-\ln p_{\hat{\theta}}\left(Y\right)\right] \\
&\overset{\mathrm{Cor}\,3.6}{=} -A^*\left(\nabla A\left(\theta\right)\right) - \mathbb{E}\left[\ln h\left(Y\right)\right] + d_A\left(\theta, \mathbb{E}\left[\hat{\theta}\right]\right) + \mathbb{B}_A\left[\hat{\theta}\right] \\
&= H\left(\mathrm{sm}\left(z\right)\right) - \mathbb{E}\left[\ln 1\right] + d_{\mathrm{LSE}}\left(z, \mathbb{E}\left[\hat{z}\right]\right) + \mathbb{B}_{\mathrm{LSE}}\left[\hat{z}\right] \\
&= H\left(Q\right) + d_{\mathrm{LSE}}\left(\mathrm{sm}^{-1}\left(Q\right), \mathbb{E}\left[\hat{z}\right]\right) + \mathbb{B}_{\mathrm{LSE}}\left[\hat{z}\right].
\end{aligned}
\tag{37}
$$

B.6 Proof of Proposition 3.8

We prove each property in the following. The arguments are constructed in a generality such that the functional case is always covered.

B.6.1 General law of total variance

Let $G\colon U \to \mathbb{R}$ be a convex function on a convex subset of a vector space. This includes the case of a vector space consisting of functions. Assume that X and Y are random variables, where X has observations in U. If $\mathbb{E}[\mathbb{E}[G(X) \mid Y]]$ exists, then by Tonelli's theorem and Jensen's inequality the other integrals in the following also exist and we have

$$
\begin{aligned}
&\mathbb{E}[\mathbb{B}_G[X \mid Y]] + \mathbb{B}_G[\mathbb{E}[X \mid Y]] \\
&= \mathbb{E}[\mathbb{E}[G(X) \mid Y] - G(\mathbb{E}[X \mid Y])] + \mathbb{E}[G(\mathbb{E}[X \mid Y])] - G(\mathbb{E}[\mathbb{E}[X \mid Y]]) \\
&= \mathbb{E}[\mathbb{E}[G(X) \mid Y]] - G(\mathbb{E}[\mathbb{E}[X \mid Y]]) \\
&= \mathbb{E}[G(X)] - G(\mathbb{E}[X]) \\
&= \mathbb{B}_G[X].
\end{aligned}
\tag{38}
$$

B.6.2 Proof of Equation 2

Let ϕ be a convex function in a vector space and $X_1, \ldots, X_{2^n}$ i.i.d. random variables such that $\mathbb{E}[\phi(X_1)]$ exists. Since $\phi\left(\mathbb{E}\left[2^{-n+1}\sum_{i=1}^{2^{n-1}} X_i\right]\right) = \phi\left(\mathbb{E}\left[2^{-n}\sum_{i=1}^{2^n} X_i\right]\right)$ due to i.i.d. assumption, we only have to show $\mathbb{E}\left[\phi\left(2^{-n}\sum_{i=1}^{2^n} X_i\right)\right] < \mathbb{E}\left[\phi\left(2^{-n+1}\sum_{i=1}^{2^{n-1}} X_i\right)\right]$. We do this by using Jensen's inequality for strict convexity:

$$
\begin{aligned}
\mathbb{E}\left[\phi\left(2^{-n}\sum_{i=1}^{2^n} X_i\right)\right] &= \mathbb{E}\left[\phi\left(\frac{1}{2}2^{-n+1}\sum_{i=1}^{2^{n-1}} X_i + \frac{1}{2}2^{-n+1}\sum_{i=2^{n-1}+1}^{2^n} X_i\right)\right] \\
&< \mathbb{E}\left[\frac{1}{2}\phi\left(2^{-n+1}\sum_{i=1}^{2^{n-1}} X_i\right) + \frac{1}{2}\phi\left(2^{-n+1}\sum_{i=2^{n-1}+1}^{2^n} X_i\right)\right] \\
&= \frac{1}{2}\mathbb{E}\left[\phi\left(2^{-n+1}\sum_{i=1}^{2^{n-1}} X_i\right)\right] + \frac{1}{2}\mathbb{E}\left[\phi\left(2^{-n+1}\sum_{i=2^{n-1}+1}^{2^n} X_i\right)\right] \\
&\stackrel{\text{iid}}{=} \mathbb{E}\left[\phi\left(2^{-n+1}\sum_{i=1}^{2^{n-1}} X_i\right)\right].
\end{aligned}
\tag{39}
$$

In combination with the definition of Bregman Information follows the statement in Equation 2.

B.6.3 Limit case

Let ϕ and $X_1, \ldots, X_n$ be defined as in the previous proof with finite mean $\mathbb{E}[X_1]$. Additionally, ϕ is almost surely continuous. Due to the definition of Bregman Information, we only have to show that

$$
\lim_{n\to\infty} \phi\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \stackrel{\text{a.s.}}{=} \phi(\mathbb{E}[X_1]).
\tag{40}
$$

Theorem 8.32 in (Capiński & Kopp, 2004) gives $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i \stackrel{\text{a.s.}}{=} \mathbb{E}[X_1]$. Note that we have in general for any random variable X with finite mean that $\{\omega \in \Omega \mid X(\omega) = \mathbb{E}[X]\} \subset \{\omega \in \Omega \mid \phi(X(\omega)) = \phi(\mathbb{E}[X])\}$. It follows with the initial
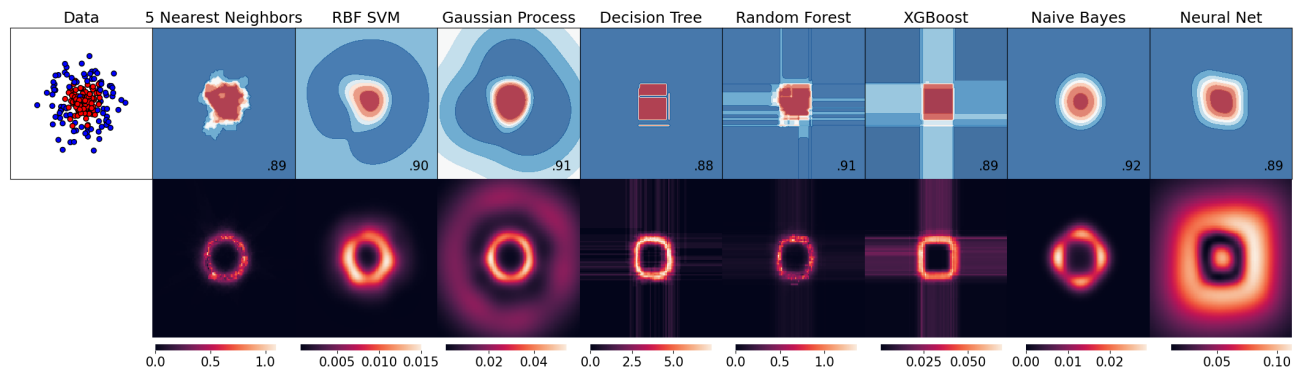
Figure 8: **Top**: Several classifiers are trained on a simulated circular task and their predictions are shown around the input space. The number in the bottom right corner is the accuracy. **Bottom**: The Bregman Information of these classifiers is estimated based on several training runs for the identical space.

conditions that

$$
\begin{aligned}
1 &= \mathbb{P}\left(\left\{\omega \in \Omega \mid \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i(\omega) = \mathbb{E}[X_1]\right\}\right) \\
&\leq \mathbb{P}\left(\left\{\omega \in \Omega \mid \phi\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i(\omega)\right) = \phi\left(\mathbb{E}[X_1]\right)\right\}\right) \\
&= \mathbb{P}\left(\left\{\omega \in \Omega \mid \lim_{n\to\infty} \phi\left(\frac{1}{n}\sum_{i=1}^{n} X_i(\omega)\right) = \phi\left(\mathbb{E}[X_1]\right)\right\}\right) \leq 1.
\end{aligned}
\tag{41}
$$

Consequently, Equation (40) holds and with it the statement $\lim_{n\to\infty} \mathbb{B}_\phi\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \overset{\text{a.s.}}{=} 0$.

C EXTENDED EXPERIMENTS

In this section, we give additional details to the experiments in the main paper, and also provide further results of extended experiments. In Section C.1, we conduct additional simulation studies to compare common classifiers in terms of their Bregman Information similar to Figure 5 and 10. Further, we investigate our proposed Bregman Information threshold algorithm in more detail on CIFAR-10 (-C) and ImageNet (-C) in Section C.2. We also showcase even stronger performance gains of our approach when using the negative log-likelihood for comparison instead of the classification accuracy.

C.1 Simulations of Toy Tasks for Common Classifiers

As already mentioned in Section 4, we compare a neural network with the classifiers k-nearest neighbors, Support Vector Machine, Decision Tree, Random Forest, XGBoost, Naive Bayes, and a neural network. The neural network is implemented via PyTorch (Paszke et al., 2019). It has a single hidden layer and 100 nodes. It is trained with the log-likelihood as criterion, the Adam optimizer provided by PyTorch, and early stopping (we split off 30% of the training set). For the other classifiers, we use the implementations from Scikit-Learn (Pedregosa et al., 2011). The hyperparameters are the following. The k-nearest neighbors uses $k = 5$, the SVM classifier uses $C = 1$ and $\gamma = 2$, the gaussian process classifier uses the RBF kernel. For Random Forests and XGBoost, we use an ensemble size of ten. The naive bayes classifer uses a gaussian assumption. All the other hyperparameters are defaults by Scikit-Learn.

The simulated data sets have 300 train instances, and 200 test instances. We construct two more toy tasks: One of circular shape with closed decision boundary, the other of linear shape. The results are depicted in Figure 8 and 9. The Bregman Information in the main paper and these figures are based on 64 training set samples. As can be seen, SVMs and Gaussian
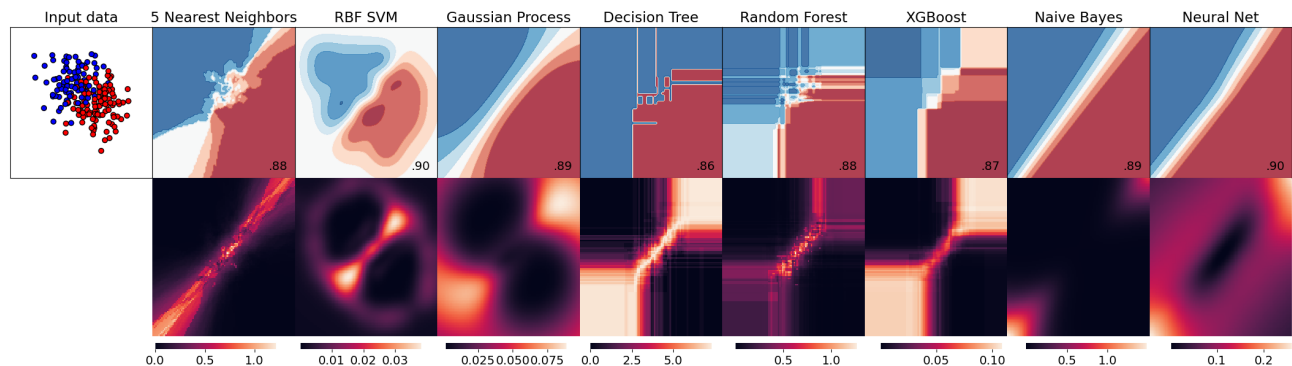
Figure 9: **Top**: Several classifiers are trained on a simulated linear task and their predictions are shown around the input space. The number in the bottom right corner is the accuracy. **Bottom**: The Bregman Information of these classifiers is estimated based on several training runs for the identical space.

Processes can indicate where the training distribution ends, while the BI of other classifiers such as KNN only identifies the direction of the decision boundary. This might make SVMs and Gaussian Processes a potential tool for out-of-domain detection for low-dimensional data.

Surprisingly, the neural network shows its lowest uncertainty around the decision boundary. Even in areas, where are sufficiently enough data samples of a class, the neural network shows uncertainty where other classifiers do not. We hypothesis a possible reason for this might be that neural networks are optimized via gradient descent and the log-likelihood, which requires anchor points of both classes for a stable convergence. Around areas with instances of only a single class, gradient descent is missing an anchor and does not 'know' how far to fit the model towards this class. At first, this might discourage using Bregman Information for out-of-domain detection at high-dimensional tasks, such as image data, fitted with a neural network. But, the traversing of the decision boundary from in-domain to out-of-domain still gives gives the highest Bregman Information of the neural network in our simulations. Consequently, in the high-dimensional setting, if most data instances lie on the decision boundary and the decision boundary is 'open' in a variety of directions, we might still receive sufficient indication of in- and out-of-domain areas in the input space. Our results in Section 4 and Section C.2 support this hypothesis.

Similar to Figure 6, we provide the same approximations and MC Dropout for additional toy tasks in Figure 10. In all cases, for the Deep Ensemble (Lakshminarayanan et al., 2017) we use 64 models, for MC Dropout (Gal & Ghahramani, 2016) an ensemble size of 5000, and for the 'real' BI we use 64 training set samples. Again, the results in Section 4 and Section C.2 support that the low-dimensional findings hold to some degree for real-world image data.

C.2   Additional Out-of-Distribution Results on CIFAR-10 and ImageNet, and Further Details

In this section, we provide further results for uncertainty thresholds in the out-of-distribution setting of CIFAR-10 (Krizhevsky, 2009) and ImageNet (Krizhevsky, 2009). We will also discuss further experiment details.

Comparisons via negative log-likelihood instead of accuracy   The log-likelihood is a proper score and as such a measure of predictive uncertainty. It captures the correctness of a predicted probability instead of only the correctness of the predicted class, like accuracy. Consequently, the log-likelihood indicates how trustworthy confidence scores are. We conduct similar experiments as in the main paper but replace the accuracy with the log-likelihood. The results can be seen in Figure 11. The performance improvement of Bregman Information with Deep Ensembles for out-of-domain instances is substantial compared to Confidence scores.

Datasets   To compare in-domain with out-of-domain performance, we use corrupted versions of the test sets introduced in (Hendrycks & Dietterich, 2019). The test sets CIFAR-10-C and ImageNet-C have 5 different severities for 20 different corruptions: Brightness, fog, glass blur, pixelate, spatter, contrast, frost, impulse noise, saturate, speckle noise, defocus
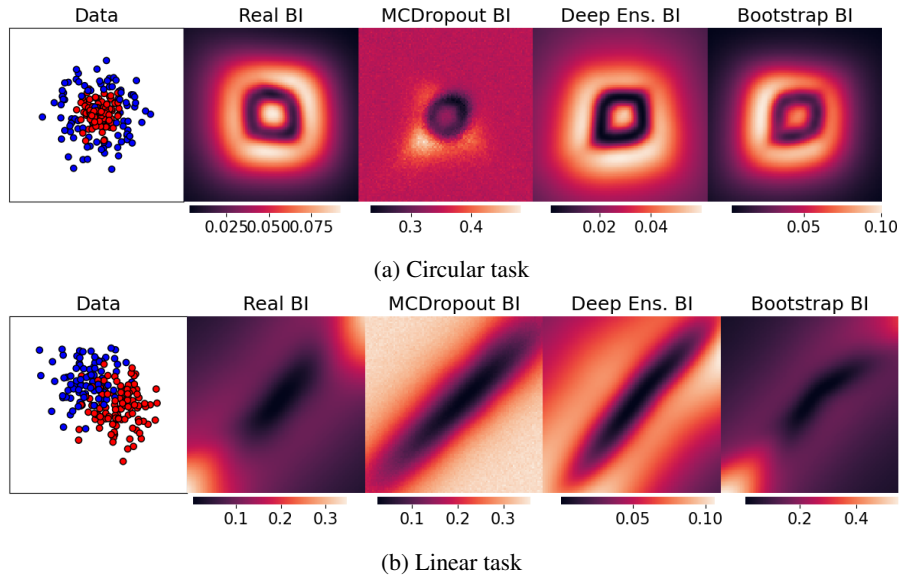
(a) Circular task

(b) Linear task

Figure 10: Different approximations of the Bregman Information for a neural network. 'Real BI' refers to the approximation via training set samples from the real distribution. The other approximations are only with respect to a single training set.

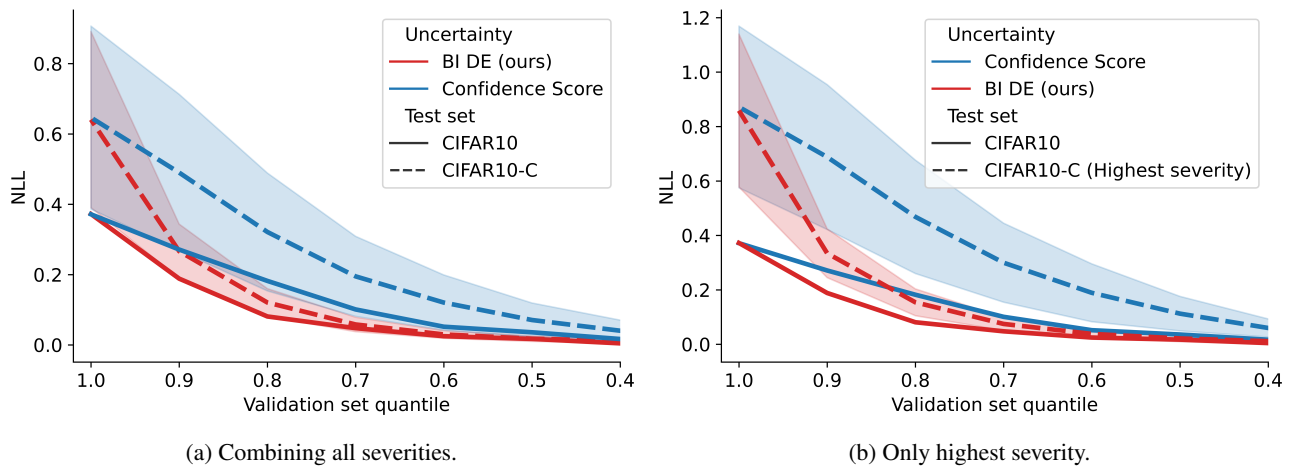(a) Combining all severities.

(b) Only highest severity.

Figure 11: Negative Log-Likelihood after discarding test instances with high levels of uncertainty for CIFAR-10 and CIFAR-10-C. Fewer samples have to be discarded to reach better NLL when using the Bregman Information as uncertainty measure.

(a) Deep Ensembles (all Severities)

(b) Test-Time Augmentation (all Severities)

(c) Deep Ensembles (only Severity 5)
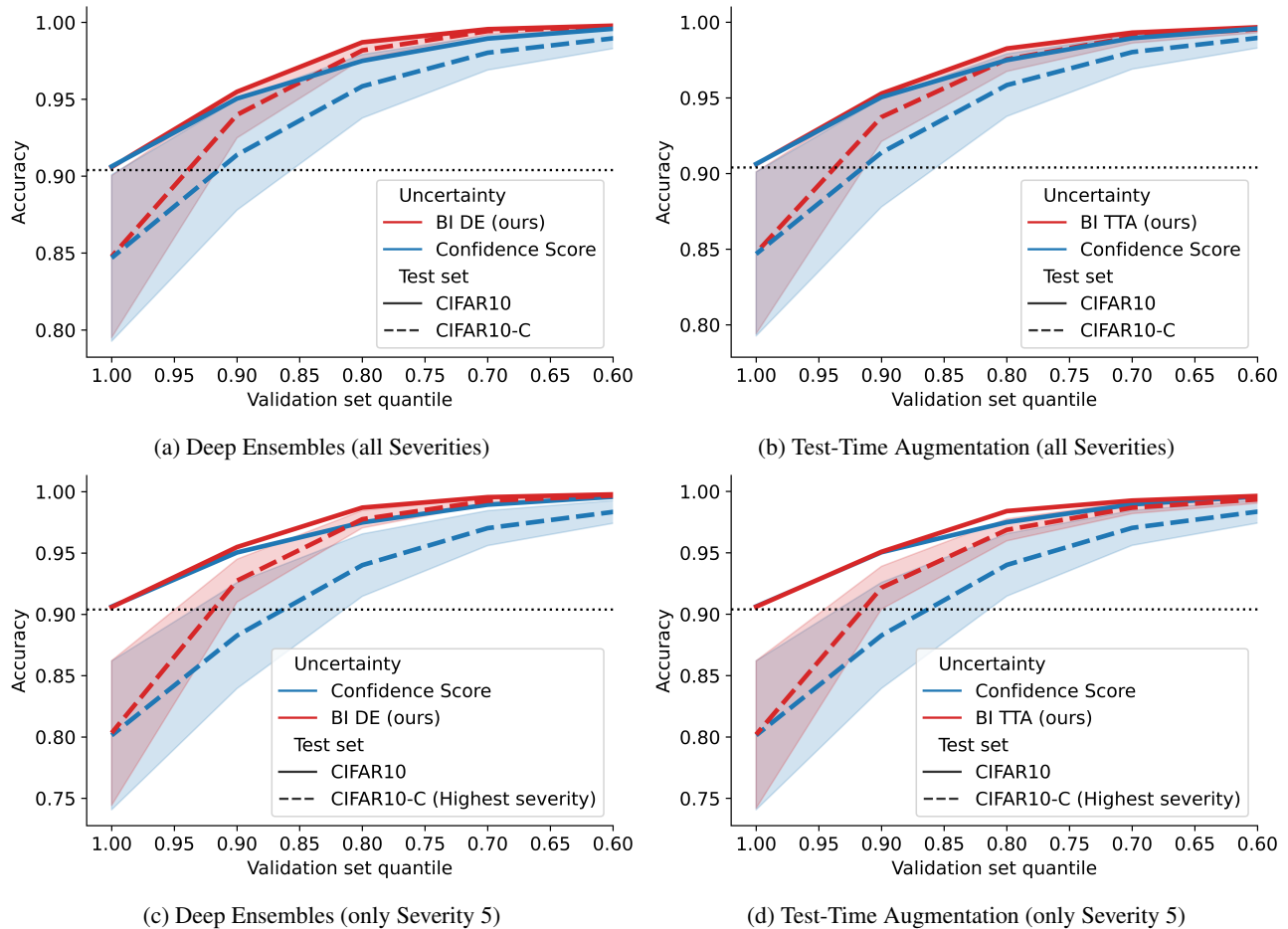
(d) Test-Time Augmentation (only Severity 5)

Figure 12: Accuracy after discarding test instances with high levels of uncertainty for CIFAR-10 and CIFAR-10-C. Fewer samples have to be discarded to reach better accuracy when using the Bregman Information as uncertainty measure.

blur, gaussian blur, jpeg compression, shot noise, zoom blur, elastic transform, gaussian noise, motion blur, and snow. For CIFAR-C-10, we have 10000 test instances per corruption per severity. We have to remove 10000 instances in each ImageNet-C corruption severity, which are corruptions of our validation set, leaving us 40000 test instances per corruption per severity.

Models and ensembles   For classification on CIFAR-10, we use a ResNet20 trained with Adam and early stopping (He et al., 2016) based on the PyTorch framework. We train the Deep Ensemble of size 10 by training the same architecture with different weight initializations.

For ImageNet, we use ResNet50 models downloaded from (Ashukha et al., 2020).[1] We also use an Deep Ensemble of size 10.

Further, we use the ensembling technique Test-Time Augmentation (Wang et al., 2019). The augmentations are Random Crop, Random Flip, and we use an ensemble size of 20.

Figure 12 shows similar results for TTA as Figure 1 in the main paper. Further, our approach still dominates when we include all corruption severities. Note that the deviation bounds are smaller, which indicates that BI is more robust for different types of corruptions.

Uncertainty threshold algorithm for Confidence scores   We also provide the Algorithm 1 adjusted to Confidence scores. It is described in Algorithm 2. The only difference is that we are not using an ensemble anymore and we flip the threshold,

---

[1]https://github.com/SamsungLabs/pytorch-ensembles

since higher confidence means less uncertainty, while higher BI means lower uncertainty.

Algorithm 2 Classifying with uncertainty threshold via Confidence scores.

Require: Validation set $\mathscr{D}$, model, $q \in [0, 1]$, test instance x'
    ConfScores $\leftarrow [\max_i \text{model}(x)_i$ for $x \in \mathscr{D}]$ $\triangleright$ Highest predicted probability for each instance
    threshold $\leftarrow$ quantile(ConfScores, q)
    if $\max_i \text{model}(x')_i <$ threshold **then**
        label as OOD $\triangleright$ Warning in real-world application
    else
        return model(x')
    end if