A Bias-Variance Decomposition for Bayesian Deep Learning

James Brofos[*]
Yale University

Rui Shu
Stanford University

Roy R. Lederman
Yale University

Abstract

We exhibit a decomposition of the Kullback-Leibler divergence into terms corresponding to bias, variance, and irreducible error. Our particular focus in this work is Bayesian deep learning and in this domain we illustrate the application of this decomposition to adversarial example identification, to image segmentation, and to malware detection. We empirically demonstrate qualitative similarities between the variance decomposition and mutual information.

1 Introduction

Bias-variance decompositions have an important role in the theory and practice of statistics. They can be used to understand the behavior of estimators, to deduce optimality criteria, and to diagnose systemic abnormalities arising from under- and over-fitting. Bias-variance decompositions can also be leveraged to characterize unusual observations based on their uncertainty characterization. Perhaps the most well-known bias-variance decomposition is the following.

Example 1.1 (Mean Squared Error Decomposition)**.** Let $y = \bar{y} + \epsilon$ where $\mathbb{E}\epsilon = 0$ and $\text{Var}(\epsilon) = \sigma^2$ and $\bar{y} \in \mathbb{R}$. Let $\hat{y}_\theta$ be a predictor of y, with the parameters of the predictor distributed as $\theta \sim \pi$. We note that this characterization comprises a broad class of inference problems, including the case where y and $\hat{y}_\theta$ may depend additionally on a set of features as in regression problems. In this case, θ can be the weights of a neural network and the predictor would also be a function of the input to the network.

Then the mean squared error of $\hat{y}$ and y decomposes as,

$$\mathop{\mathbb{E}}_{\epsilon}\mathop{\mathbb{E}}_{\theta}(y - \hat{y}_\theta)^2 = (\bar{y} - \mathop{\mathbb{E}}_{\theta}\hat{y}_\theta)^2 + \text{Var}(\hat{y}_\theta) + \sigma^2 \tag{1}$$

This first term is the squared "bias" between the true mean and the predictive mean, the second is the variance of the prediction, and the third is the irreducible error (noise that is inherent to the generative process for y that could never be eliminated). Notice that in this decomposition there is a natural sense in which $\tilde{y} = \mathbb{E}_\theta \hat{y}_\theta$ is a "centroid" of the random variable $\hat{y}_\theta$ for a squared error loss function; it is well-known that the expectation of a random variable minimizes the squared distance of a random draw to a fixed point. Therefore, the bias captures the squared distance of the true mean to the centroid while the variance captures the average squared distance of a random $\hat{y}_\theta$ to $\tilde{y}$. This notion of a centroid will be relevant to our decomposition.

We consider the intersection of the bias-variance decomposition with the field of Bayesian deep learning. Although the bias-variance decomposition contained in this work is not specific to Bayesian deep learning in particular, we believe that there is an opportunity to exhibit the benefits of a bias-variance decomposition to applications involving Bayesian neural networks. In contrast to point-estimate neural network models, the promise of Bayesian deep learning is to provide uncertainty

[*]Approved for Public Release; Distribution Unlimited. Public Release Case Number 19-2139.

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.

estimates for parameters. The principal difficulty in utilizing Bayesian methods in deep learning is the intractability of the posterior; therefore, we leverage a form of variational inference in order to efficiently generate approximate samples from the neural network posterior. In this work, we seek to propagate this uncertainty characterization through a decision-theoretic lens in order to obtain a general bias-variance decomposition for Bayesian deep learning. An objective in the empirical assessment portion of this work is to assess the usefulness of the bias-variance decomposition on a variety of compelling and real-world applications.

The organization of the paper is as follows. In section 3 we first establish the objective and methods of Bayesian deep learning. Thereafter, we review the pertinent theory and describe, using information-theoretic techniques, a bias-variance characterization that is suitable for Bayesian deep learning, with particular focus on classification tasks. This section includes multiple examples that demonstrate the character of proposed decomposition. Section 2 examines related research on general bias-variance decompositions. In section 4 we examine the application of the bias-variance decomposition in several domains: (i) we demonstrate the usefulness of the decomposition in for characterizing unusual and adversarial inputs; (ii) we apply the bias-variance decomposition to a per-pixel image segmentation task and show the effect of contraction of the posterior on the decomposition; (iii) we utilize the bias-variance decomposition to examine software executables and illustrate the regimes of high-uncertainty predictions and evasive malware. In section 5 we conclude and offer some perspective on future work.

2 Related Work

The most general treatment of bias-variance decompositions is [8], which considers Bregman divergences. Such a general approach is appealing, but suffers the immediate drawback that a notion of a "centroid" for the Bregman divergence may not have a convenient form. [5] also develops a bias-variance decomposition for general loss functions, but breaks with the tradition that loss should equal the sum of bias and variance terms and specializes to classification tasks where the loss is computed as a function of classification decisions (in contrast, we will adopt a distribution-theoretic approach by seeking to estimate a distribution over classes). The bias-variance decomposition we derive in corollary 3.2 arises as a special case of theorem 3.1; however, the corollary 3.2 has been derived directly in [2]. We see an opportunity for a rigorous empirical evaluation of this bias-variance decomposition, which is the primary purpose of this work.

3 Theory

In this section we introduce theory for Bayesian deep learning and the information-theoretic treatment of the bias-variance decomposition. In section 4 we combine these frameworks in application to several domains.

3.1 Bayesian Deep Learning

Bayesian deep learning offers a probabilistic alternative to neural networks estimated via maximum a posteriori or maximum likelihood. Let θ denote the parameters of a neural network; given a prior $\pi(\theta)$, Bayesian deep learning infers a posterior

$$\pi(\theta|\mathbf{Y}, \mathbf{X}) \propto L(\mathbf{Y}|\mathbf{X}, \theta)\pi(\theta) \tag{2}$$

where L is the likelihood function. This work focuses on classification tasks and takes L to be the negative categorical cross entropy. For regression tasks a natural choice of likelihood function is the exponentiated mean squared error.

With the Bayesian neural network posterior we can compute the posterior predictive distribution of $\mathbf{y}^*$ given $\mathbf{x}^*$:

$$f(\mathbf{y}^*|\mathbf{x}^*) = \int f_\theta(\mathbf{y}^*|\mathbf{x}^*)\pi(\theta|\mathbf{Y}, \mathbf{X})\, \mathrm{d}\theta \tag{3}$$

This has the effect of integrating over uncertainty in the network weights. The principle difficulty in Bayesian deep learning, as in other Bayesian methods, is that the normalizing constant of posterior

is intractable, requiring an $|\theta|$-dimensional integral over the parameter space. In the numerical examples in this work, we adopt an approximate inference procedure called Monte Carlo dropout. This procedure estimates the true posterior by multiplying the per-layer weight matrices by diagonal matrices consisting of i.i.d. Bernoulli random variables with *failure* probability p.

3.2 A Bias-Variance Decomposition

In this work we will use $D(P\|Q)$ to denote the KL-divergence between probability measures P and Q. We use the notation f_θ to denote a probability distribution depending on (possibly random) parameters θ.

Theorem 3.1 (Bias-Variance-Irreducible Error Decomposition of the KL-Divergence)**.** *Let $\theta \sim \pi$ and $\omega \sim \pi'$ and let f_θ be a distribution with the same support as a random probability distribution P_ω for all θ and ω. Then, if $Y \sim \bar{P}$,*

$$\mathop{\mathbb{E}}_{\omega\sim\pi'}\mathop{\mathbb{E}}_{\theta\sim\pi} D(P_\omega\|f_\theta) = \underbrace{D(\bar{P}\|\tilde{P})}_{\text{“squared bias”}} + \underbrace{\mathop{\mathbb{E}}_{\theta\sim\pi} D(\tilde{P}\|f_\theta)}_{\text{“variance”}} + \underbrace{I(Y;\omega)}_{\text{“irreducible error”}} \tag{4}$$

where $\bar{P}$ is the marginal

$$\bar{P}(\cdot) = \mathop{\mathbb{E}}_{\omega\sim\pi'} P_\omega(\cdot) \tag{5}$$

and where

$$\tilde{P}(\cdot) \propto \exp\left(\mathop{\mathbb{E}}_{\theta\sim\pi} \log f_\theta(\cdot)\right) \tag{6}$$

is the “geometric mixture” of f_θ over $\theta \sim \pi$.

The proof of theorem 3.1 relies on the “forward” and “backward” compensation identities in information theory; see appendix B. The first term of the decomposition is naturally regarded as (squared) bias because it measures the closeness, in terms of KL-divergence, of the target distribution P to the geometric mixture; theorem 3.4 shows why the geometric mixture has a meaningful interpretation as a centroid of the posterior. The second term in the decomposition has an interpretation as the variance because it measures the average divergence of a random f_θ to the geometric mixture centroid. Under this interpretation, both the bias and the variance are measured in units of the logarithmic base, which in these experiments are natural units (nats).

An important special case of theorem 3.1 is when $P \equiv P_\omega$ so that there is no randomness in the first argument to the LHS KL-divergence.

Corollary 3.2 (Brinda, Klusowski, and Yang [2])**.** *Suppose, in addition, that $P \equiv P_\omega$. Then, there is no irreducible error and the bias-variance decomposition becomes,*

$$\mathop{\mathbb{E}}_{\theta\sim\pi} D(P\|f_\theta) = D(P\|\tilde{P}) + \mathop{\mathbb{E}}_{\theta\sim\pi} D(\tilde{P}\|f_\theta) \tag{7}$$

Corollary 3.2 can be derived directly from the reverse compensation identity.

3.3 The Geometric Mixture Centroid

The geometric mixture $\tilde{P}$ does not feature prominently in machine learning so it is worth recalling some properties it exhibits. We begin by offering some results that motivate why one should consider the geometric mixture at all. The following results are due to [2, 3] and a proof is contained in appendix B.

Theorem 3.3 (The Geometric Mixture is a Bayes Rule)**.** *Consider the KL-divergence loss $L(\theta, P(\Lambda)) = D(P(\Lambda)\|f_\theta)$ where Λ is data and $P(\Lambda)$ is an estimator of f_θ. Then the geometric mixture $\tilde{P}(\cdot) \propto \exp\left(\mathop{\mathbb{E}}_{\theta|\Lambda} \log f_\theta(\cdot)\right)$ is the Bayes rule; this means that $\tilde{P}(\cdot)$ is the estimator that minimizes the posterior risk:*

$$\tilde{P} = \arg\min_P \mathop{\mathbb{E}}_{\theta|\Lambda} D(P\|f_\theta) \tag{8}$$

Moreover, the geometric mixture is an admissible decision rule under this loss.

Note that although this is an appealing property it is not immediately relevant to most classification tasks which, rightly, use the KL-divergence in the reverse direction. For our purposes, a better motivation for the geometric mixture is that it represents a "most representative distribution" – that is, it may be regarded as a centroid. The following result formalizes this intuition which is used in theorem 3.3.

Theorem 3.4 (The Geometric Mixture Minimizes KL-divergence)**.**

$$\tilde{P} = \arg\min_{P} \mathop{\mathbb{E}}_{\theta\sim\pi} D(P\|f_\theta) \tag{9}$$

This follows immediately from corollary 3.2 because the KL-divergence is non-negative.

Oftentimes in statistics we are dealing with distributions that are members of an exponential family. The categorical distribution is a special case that is considered in this work, but geometric mixtures exhibit special properties for the more general class of exponential family distributions and these are contained in appendix B. As this work concerns classification tasks in particular, and since the categorical distribution is a member of the exponential family, these results are relevant to our discussion here.

3.4 Illustrative Theoretical Examples

One may wonder if the geometric mixture could be substituted for the usual marginal distribution $\hat{P}(y) = \mathbb{E}_\theta f_\theta(y)$ (or "arithmetic mixture") in prediction tasks. Although in practice the prediction accuracies are similar, the KL-divergence loss in the direction $D(P\|\hat{P})$ is typically smaller than $D(P\|\tilde{P})$. This leads us to an important recommendation that while $\tilde{P}$ is useful for decomposition of the expected KL-divergence loss it is not optimal for minimizing prediction loss. Indeed, $\hat{P}$ can be shown to be the Bayes rule for the loss function $D(f_\theta\|P(X))$; see [2] for a proof of this using theorem 3.1.

Example 3.1 (Single Bernoulli Trial)**.** Suppose we adopt a generative model

$$Y|\theta \sim \text{Bern}(\theta) \tag{10}$$

$$\theta \sim \text{Unif}(0,1) \tag{11}$$

Let us suppose also for the sake of argument that there exists a true θ^* from which a single observation $Y = 1$ is generated. The posterior distribution is therefore,

$$\pi(\theta|Y=1) = 2\theta \tag{12}$$

on the support $\theta \in (0,1)$. Under this posterior the arithmetic and geometric mixtures are seen to be,

$$\hat{P}\left[Y=1\right] = \frac{2}{3} \tag{13}$$

$$\tilde{P}\left[Y=1\right] = \frac{\exp(-1/2)}{\exp(-1/2)+\exp(-3/2)} \tag{14}$$

Taking $\pi(\theta|Y=1)$ as the believed distribution, we can decompose its expected KL-divergence loss as

$$\mathop{\mathbb{E}}_{\theta\sim\pi(\theta|Y=1)} D(\theta^*\|\theta) = \underbrace{\theta^*\log\frac{\theta^*}{\tilde{P}(1)} + (1-\theta^*)\log\frac{1-\theta^*}{\tilde{P}(0)}}_{\text{"squared bias"}} + \underbrace{e^{-1/2}+e^{-3/2}}_{\text{"variance"}} \tag{15}$$

Finally, to reinforce the original point that the geometric mixture is useful as a tool for decomposition and not for prediction, consider the following exact calculations of the Bayes risk.

$$\mathop{\mathbb{E}}_{\theta,Y} D(\theta\|\hat{P}) = \frac{2}{9}(\log\frac{27}{8} - 1) + \frac{\log 729 - 5}{18} \tag{16}$$

$$\approx 0.136514 \tag{17}$$

$$\mathop{\mathbb{E}}_{\theta,Y} D(\theta\|\tilde{P}) = 3\log(1+e) - 4 + \frac{6\log(1+e)-5}{18} \tag{18}$$

$$\approx 0.146595 \tag{19}$$

One may be inclined to think that the variance term could roughly correspond with entropy. However, uncertainty about the distribution is not equivalent to uncertainty about an outcome generated by that distribution. The following two examples demonstrate that these two quantities may not necessarily coincide.

Example 3.2 ($\tilde{P}$ can have high entropy and zero variance.)**.** Let $f \equiv f_\theta$ be a uniform p.m.f. on k classes. Then $\tilde{P} = f$ and $H(\tilde{P}) = \log k$ is maximum entropy. Nonetheless,

$$\begin{aligned} \mathop{\mathbb{E}}_{\theta\sim\pi} D(\tilde{P}\|f_\theta) &= \mathop{\mathbb{E}}_{\theta\sim\pi} D(f\|f_\theta) && (20)\\ &= \mathop{\mathbb{E}}_{\theta\sim\pi} D(f\|f) && (21)\\ &= 0 && (22) \end{aligned}$$

Example 3.3 ($\tilde{P}$ can have low entropy and high variance.)**.** Let $\theta \in (\theta_1, \theta_2)$ index two distributions supported on $\{0, 1\}$. Let $f_{\theta_1}(0) = 1 - \delta$ and $f_{\theta_2}(0) = \delta$ and consider a $\pi(\theta)$ that places $1 - \epsilon$ probability on θ_1 and ϵ probability on θ_2, for $\delta, \epsilon \in (0, 1)$. Then

$$\mathop{\mathbb{E}}_{\theta\sim\pi} D(\tilde{P}\|f_\theta) = -\log\left(e^{\epsilon\log(1-\delta)+(1-\epsilon)\log\delta} + e^{\epsilon\log\delta+(1-\epsilon)\log(1-\delta)}\right) \tag{23}$$

Suppose $\epsilon > 1/2$ and let (δ_n) be a sequence converging to zero. Then $\tilde{P}_n(0) \to 1$ and $\mathop{\mathbb{E}}_{\theta\sim\pi} D(\tilde{P}_n\|f_\theta)$ is unbounded. Also observe that under these same conditions the marginal satisfies $\hat{P}(0) \to 1 - \epsilon$; this shows a case in which the marginal and geometric mixture differ substantially.

One notices that the variance term in the decomposition bears some superficial resemblance to the mutual information between $Y \sim \mathbb{E}_\theta f_\theta$ and θ. Indeed, recall the characterization of mutual information as

$$I(Y;\theta) = \mathop{\mathbb{E}}_{\theta\sim\pi} D(f_\theta\|\mathbb{E}_\theta f_\theta) \tag{24}$$

In contrast the variance component reverses the position of f_θ and substitutes the marginal distribution of Y for its geometric mixture. However, the "structural similarity" between the mutual information and the variance component suggests a connection which will be explored empirically. Indeed, one appealing point of comparison is the interpretation of mutual information as the "epistemic uncertainty" – or uncertainty due to a dearth of training data.

4 Applications

In these experiments consider $\theta \sim \pi$ to be an approximate draw from the posterior over the parameters of the neural network. We use Monte Carlo dropout to collect n i.i.d. approximate samples from the posterior, denoted $\{\theta_i\}_{i=1}^n$. Let f_{θ_i} represent the distribution over classes produced by a particular sampled parameter vector θ; the fact that f_{θ_i} could depend on an addition input $\mathbf{x}$ (e.g. an input image) is suppressed for notational brevity. Every term in the bias-variance decomposition, in addition to information-theoretic quantities, may be approximated.

$$\tilde{P}_{\text{approx.}}(\cdot) \overset{\text{def.}}{\propto} \exp\left(\frac{1}{n}\sum_{i=1}^n \log f_{\theta_i}(\cdot)\right) \tag{25}$$

$$\hat{P}_{\text{approx.}}(\cdot) \overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n f_{\theta_i}(\cdot) \tag{26}$$

$$\mathop{\mathbb{E}}_{\theta\sim\pi} D(\tilde{P}\|f_\theta) \approx \frac{1}{n}\sum_{i=1}^n D(\tilde{P}_{\text{approx.}}\|f_{\theta_i}) \tag{27}$$

$$D(P\|\tilde{P}) \approx D(P\|\tilde{P}_{\text{approx.}}) \tag{28}$$

$$H(\hat{P}) \approx H(\hat{P}_{\text{approx.}}) \tag{29}$$

$$I(Y;\theta) \approx H(\hat{P}_{\text{approx.}}) - \frac{1}{n}\sum_{i=1}^n H(f_{\theta_i}) \tag{30}$$

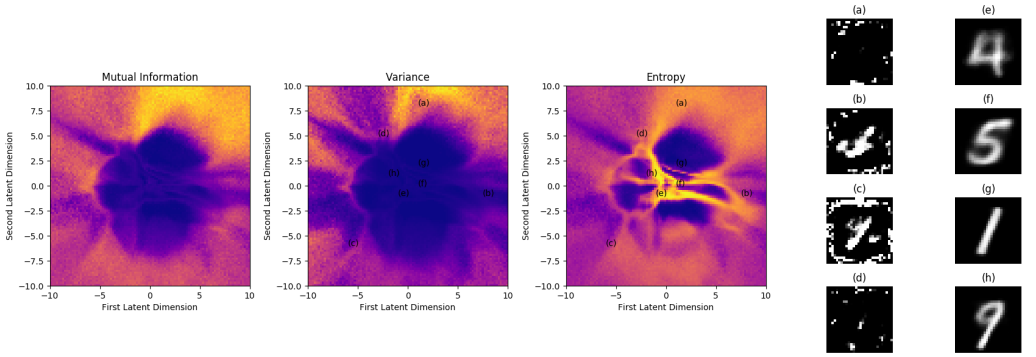

Figure 1: Visualization of the latent space uncertainty estimated by a variational autoencoder. The variance component, mutual information, and predictive entropy are compared; notice in particular that variance is low in the center region despite the presence of decision boundaries between digit classes. Decoded images from several locations of interest in the latent space are also shown. For locations near the data-dense central region, the decodings are reasonable representations of hand-written digits. However, for latent points far from any training data the decoded images are unrecognizable. For many of these bogus decoded images, both the posterior entropy and variance are large, but this is not always the case, indicating a need for more sophisticated posterior inference strategies than dropout Monte Carlo.

The main purpose of this section is to provide a detailed examination of the bias-variance decomposition for modern machine learning applications.

An important consideration in these experiments is whether or not the one-hot encoded target vector represents a worthwhile approximation to true P. After all, it is typically the case that a true probability vector over the various classes is not available. In many applications, we believe the assignment definitively to one of the classes is reasonable. Take for instance the image segmentation problem: if one observes the image, for most of the objects close to the camera you can see with certainty that the object represents a portion of a street sign, a part of a car, a pedestrian, the facade of a building, etc. This breaks down at a distance and for objects in shadow (e.g. underneath a car), though, and a higher resolution image could alleviate the problem. For our malware detection application, most varieties of malicious code do not try to obscure their destructive nature. Hence, we assert that the one-hot approximation of true P is reasonable in this application area as well. Nevertheless, at the end of the day, what we have available to us is the one-hot encoding; hence, one can regard these experiments as a "what-if" experiment to assess the bias-variance decomposition using the one-hot approximation, even if one does not fully subscribe to the belief that it is an appropriate reflection of true P.

4.1 Characterizing High Bias and High Variance Predictions

To provide intuition for the behavior of the bias-variance decomposition, we first consider the classification of hand-written digits using the MNIST dataset. All experiments using MNIST use a convolutional network architecture; after the convolutional layers an initial dropout layer with $p_{\text{drop}} = 0.25$ is applied, which is followed by a fully-connected layer with 128 hidden units and dropout probability $p_{\text{drop}} = 0.5$. We use the Adam optimizer with a learning rate of 0.001; we collect $n = 100$ i.i.d. posterior samples using Monte Carlo dropout. Our first result shows the top-five high bias and variance images in the left panel of fig. 2. High bias images are misclassified with large posterior probabilities, but observe that in several cases the predicted label of the digit is quite believable (if not believable with probability one). The images corresponding to large variances are those which exhibit substantive multimodalities in the posterior predictive distribution.

In order to investigate the usefulness of the variance component in particular, we draw inspiration from [**?**] and train a variational autoencoder with a two-dimensional latent space. For each point in the latent space, we use the variational autoencoder to decode it to an image representation. We can then compute its variance, which, importantly, does not require a true label. Each point is then colored according to the value of the variance; the same procedure is applied to the posterior predictive entropy

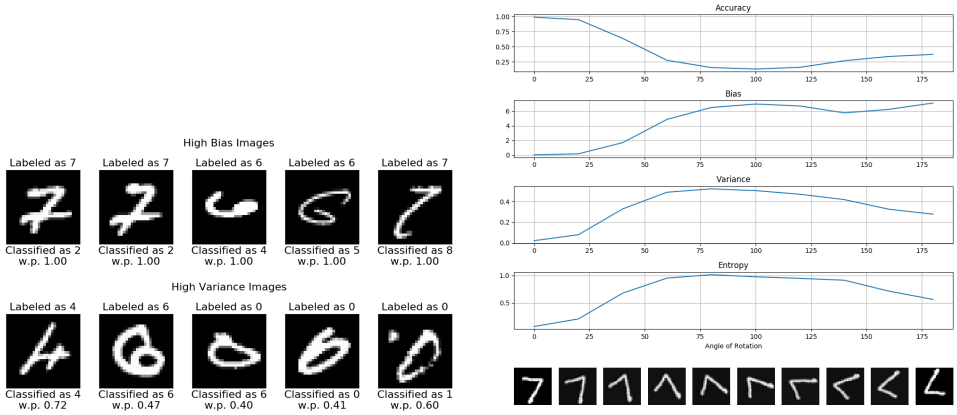

Figure 2: (Left) Hand-written images showing high-bias and high-variance samples. The high-bias imagery are those whose geometric mixture probabilities are equal to one in an incorrect class. High-variance images are ones for which the posterior samples exhibit are large amount of divergence relative to the geometric mixture centroid. (Right) Behavior of accuracy, bias, variance, and entropy when out-of-distribution images are generated by rotation. Observe that entropy and variance behave similarly under these conditions. The bias component of the decomposition reveals over-confidence in certain digit classes for large rotations despite an increasing accuracy.

of the image. An important question is "What represents a useful uncertainty characterization?" One answer is that it is desirable to have low uncertainty in data-rich regions of the latent space and high uncertainty between image classes and in regions of the latent space that are far from any training data; this is captured by the posterior predictive entropy. On the other hand, it may be desirable for the uncertainty characterization to recognize the normalcy of high entropy at decision boundaries and to only exhibit high uncertainty when there is disagreement between posterior samples; the variance component better reflects this interpretation of uncertainty characterization. Notice that consistent with [**?**], there are spurious regions of low uncertainty irrespective of whether or not entropy or variance is used; this occurs because Monte Carlo dropout is a coarse approximation of the true posterior. These spurious regions can be corrected by adopting a less crude variational family or by utilizing other sampling procedures with greater fidelity to the posterior. These results are visualized in fig. 1.

Because the out-of-distribution examples produced by the variational autoencoder are unrecognizable, they cannot be assigned to a particular image class. We therefore sought to assess the bias-variance characterization derived from the posterior but for more "organic" out-of-distribution images for which assignment of a true label was conceivable. These results are shown in fig. 2 For this purpose, we consider rotations of the MNIST digits. Both the posterior entropy and the variance component are capable of recognizing the out-of-distribution source of the rotated images as demonstrated by the relatively higher uncertainty measure. The bias decomposition also reveals that for larger angles of rotation the bias increases even as the accuracy increases; this is because even though some digits (e.g. "1" and "8") are up-down symmetric, others (e.g. "6" and "9") come to resemble digits of another class. This latter case causes the KL-divergence to increase dramatically for these digit classes.

4.2 Leveraging Variance to Detect Adversarial Examples

We next consider the bias-variance decomposition in the context of adversarial attacks against the dropout network. This permits us to evaluate the applicability of the bias-variance decomposition to increasing adversarial for the simple kinds of attacks evaluated here. In particular, we utilize the Fast Gradient Sign (FGS) attack with varying step sizes to attack the dropout network; FGS is a "white-box" attack method that assumes an adversary has access to the underlying gradients of a model. Given a network input $\mathbf{x}$, adversarial examples are computed according to the update equation,

$$\mathbf{x}_{\text{adv}} = \mathbf{x} - \epsilon \operatorname{sign}\left(\max_{y} \nabla_{\mathbf{x}} \log p(y|\mathbf{x})\right) \tag{31}$$

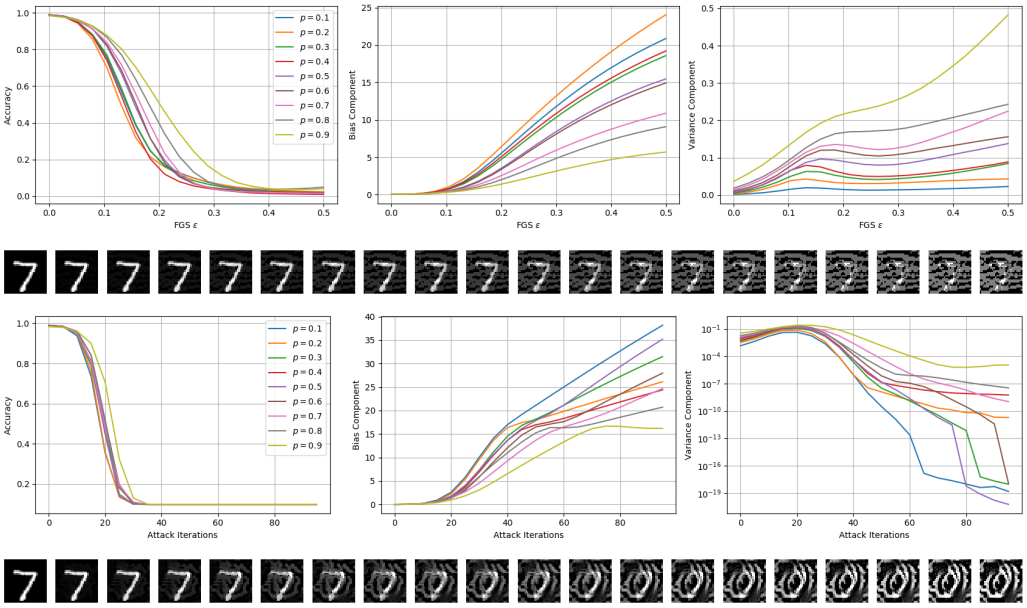

Figure 3: (Top: FGS attack.) Visualization of the bias-variance decomposition for adversarial examples. As expected, the bias component gets larger with larger values of the FGS step-size. This occurs because the adversarial examples are misclassified with higher posterior probability. The variance component also increases with FGS step-size, suggesting a mechanism to detect adversarial examples using this decomposition. Observe that higher dropout probabilities produce higher values of variance even when there is no adversarial corruption ($\epsilon = 0$); hence, any adversarial detection mechanism must be calibrated with respect to these baseline levels of uncertainty. (Bottom: Iterated FGS attack.) As in the single-step FGS attack, the accuracy and bias decrease continually as the number of attack iterations increase. The variance exhibits a different behavior, however, rising initially to reflect to abnormality of the adversarial images with a relatively small number of iterations, but subsequently decreasing to extremely small levels (note the logarithmic axis in the iterated attack).

This update has the effect of approximately reducing the the posterior probability of the most-likely class. The posterior probability $\log p(y|\mathbf{x})$ is computed using a "base network" where the network weights are scaled by a factor of $1 - p_{\text{drop}}$ so that the average activation remains the same but dropout is not actually applied. We assess the effect of varying the dropout probability p_{drop} on the uncertainty characterization. Results are shown in fig. 3. We also visualize a typical adversarial example generated by this procedure. Notice has the uncertainty characterization produced by the variance component rises significantly as the FGS step-size increases from zero. Because the variance decomposition does not require knowledge of the true class, these experiments provide preliminary evidence that the variance can detect out-of-distribution adversarial examples.

The second attack we consider is an iterative form of the FGS attack. In this case, the adversarial attacks are constructed according to the equation,

$$\mathbf{x}_{\text{adv}}^{(i+1)} = \mathbf{x}_{\text{adv}}^{(i)} + \epsilon \, \text{sign}\left(\nabla_{\mathbf{x}} \log p(y^*|\mathbf{x})\right) \tag{32}$$

where y^* is a target class for the adversarial example; this update can be understood as iteratively increasing the posterior probability of a target class. Following [7], in our experiments we fix $\epsilon = 0.01$, set the target class to "0," and evaluate the effect of an increasing number of attack iterations. Our results show that there exists an intermediate region between the initial FGS attack and the point where the classifier becomes convinced of the target class where the variance component is large relative to the baseline. With a larger number of attack iterations, the classifier becomes utterly convinced of its decision and the variance component is reduced even below the baseline level; this suggests an intriguing possibility that exceptionally *low* variance components are equally suspicious as exceptionally high ones.

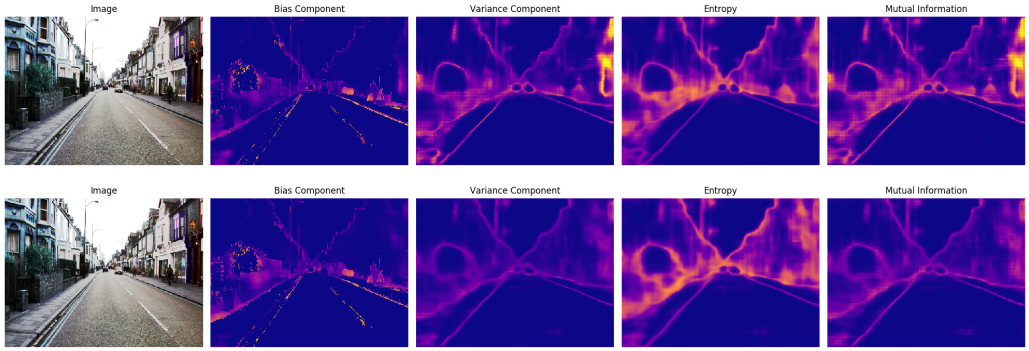

Figure 4: Test set results of performing image segmentation with Monte Carlo dropout. We show the input image, the bias and variance decomposition, the predictive entropy, and the mutual information. Notice the similarity between the variance component and the mutual information. The top row shows the results for the original training set while the bottom row shows results for the augmented dataset; observe how the additional data leads to posterior contraction, from which a reduced variance component is deduced (darker colors mean lower values of bias, variance, entropy, and mutual information). Notice that the bias component is qualitatively similar even under the increased training set, suggesting that the model is estimated to capacity.

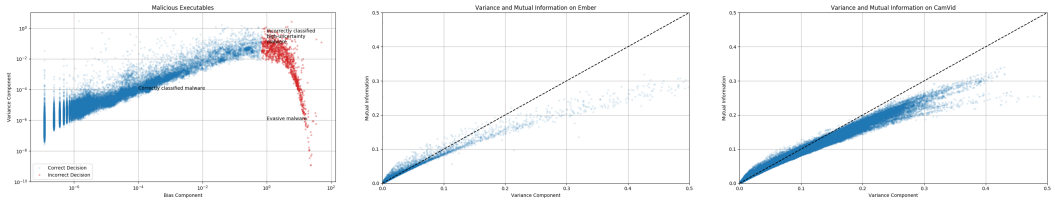

Figure 5: (Left) Bias-variance decomposition for malicious executables in the Ember dataset. Observe that the plot of bias versus variance resembles a noisy form of the plot of probability versus entropy. We use the bias-variance decomposition to characterize malware as correctly classified, high uncertainty, or evasive. (Middle) Variance component versus mutual information evaluated on the Ember malware dataset. For small values of the variance, the variance and mutual information are approximately equal; for large values of the variance, it is consistently an overestimate of the mutual information. The $x = y$-relationship is shown as a black dotted line. (Right) The variance and mutual information compared on the CamVid image segmentation task. Here again a qualitatively similar relationship between variance and mutual information appears.

4.3 Bias-Variance Decomposition in Image Segmentation

We next evaluate the bias-variance decomposition on the per-pixel image segmentation task in the autonomous driving CamVid dataset [4]. This dataset comprises 367 training images and 233 test images; the objective in this dataset is to assign every pixel in an image to one of twelve object classes such as road, building, pedestrian, or sky. The network architecture is the Bayesian SegNet introduced in [6]. Because the variance decomposition is a measure of the deviation about a centroid of the posterior, a desirable property for it to exhibit is posterior contraction. In order to evaluate this, we consider augmenting the typical training set with an additional 101 images and observe the effect on the variance. Results are visualized for a typical image in fig. 4 where comparisons to posterior predictive entropy and mutual information are also shown. Notice how the variance decreases, as desired, when additional data is used to approximate the parameter posterior of the segmentation network.

4.4 Bias-Variance Decomposition in Malware Detection

Malware detection is an important domain of computer security. To demonstrate the combination of Bayesian deep learning and the bias-variance decomposition in a domain outside of computer vision, we consider the problem of static executable analysis. In this setting, "static" features are extracted

from an executable, which are then provided as input to a machine learning system to determine if the executable has malicious functionality. This differs from dynamic analysis in which an executable payload may be detonated within a computing environment in order to collect additional salient data. Examples of features that are extracted from an executable include its imported libraries, exported functions, byte histogram data (a measure of compression), and executable header data such as linker and operating system version. In total, 2,412 features were extracted from the executables in the dataset.

In these experiments we use the Ember dataset released by Endgame Incorporated [1]. Ember comprises 600,000 training executables and 200,000 test executables, each equally partitioned between benign and malicious programs. All observations in the dataset correspond to portable executables (PE files) for Windows computers. The neural network architecture consists of a two hidden layers with 1,024 and 512 hidden neurons, respectively.

One utilization of the bias-variance decomposition is to characterize various regimes of malicious software. In fig. 5 we visualize the bias-variance decomposition for malware; because malicious programs are typically evident from their extracted features, we note that the accuracy of the model is 99.13%. The fact that executables can be classified with high confidence produces very small biases and variances in many - but not all - cases. We observe that a substantial proportion of incorrectly classified malware executables exhibit large variance, indicating that the variance decomposition can be used to identify PEs requiring additional scrutiny. There is also a regime of low-variance, high-bias malicious software: this corresponds with the notion of evasive malware, which are malicious executables that seek to conceal their destructive character by appearing benign.

5 Discussion and Conclusion

In this work we have evaluated a bias-variance decomposition suitable for Bayesian deep learning. Leveraging the notion of a "geometric mixture" we are able to decompose the expected KL-divergence into terms corresponding to the squared bias and variance of the algorithm. We provide several examples regarding this decomposition and exhibit some theory of the geometric mixture to illustrate its properties. We apply the bias-variance decomposition to situations requiring the characterization of unusual inputs, including the area of adversarial example detection and exploring the latent space of a variational autoencoder. We evaluate the bias-variance decomposition on a per-pixel image segmentation problem, where we evaluate the effect of posterior contraction on the variance component of the decomposition. Our final experiment examines the domain of malware detection, where we utilize the bias-variance decomposition to characterize regimes of malicious software. In these latter two experiments we evaluate the interesting relationship between the variance and the mutual information.

A direction for future research is the amortization of the bias-variance decomposition. The most computationally burdensome aspect of the bias-variance decomposition is the collection of multiple posterior samples, which requires n forward passes through the dropout network. It would be more efficient by far to amortize the decomposition so as to obtain the approximate bias and variance immediately.

References

[1] H. S. Anderson and P. Roth. EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models. *ArXiv e-prints*, April 2018.

[2] W.D. Brinda, Jason M. Klusowski, and Dana Yang. Hölder's identity. *Statistics & Probability Letters*, 148(C):150–154, 2019.

[3] William David Brinda. Adaptive Estimation with Gaussian Radial Basis Mixtures. 2018.

[4] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[5] Pedro Domingos. A unified bias-variance decomposition for zero-one and squared loss. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 564–569. AAAI Press, 2000.

[6] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *CoRR*, abs/1511.02680, 2015.

[7] Yingzhen Li and Yarin Gal. Dropout inference in bayesian neural networks with alpha-divergences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 2052–2061. JMLR.org, 2017.

[8] David Pfau. A Generalized Bias-Variance Decomposition for Bregman Divergences. 2013.

A Preliminaries

The proof relies on the compensation identity and reverse compensation identity from information theory. These descriptions are taken from [2, 3].

Theorem A.1 (The compensation identity)**.** *Let $\{f_\theta : \theta \in \mathcal{T}\}$ be a family of probability densities with respect to a σ-finite measure α, and suppose that $F : (\theta, y) \to f_\theta(y)$ is product measurable. Let $\Theta \sim Q$ be a $\mathcal{T}$-valued random element. Let P be probability measure defined with respect to α. Then,*

$$\mathbb{E}_{\Theta\sim Q} D(f_\Theta \| P) = D(\bar{F}_Q \| P) + \mathbb{E}_{\theta\sim Q} D(f_\theta \| \bar{F}_Q) \tag{33}$$

where $\bar{F}_Q = \mathbb{E}_\Theta f_\Theta$, a distribution over $\mathcal{Y}$ is the marginal over $\Theta \sim Q$.

Definition A.1 (Q-geometric mixture of $\{f_\theta\}$)**.**

$$\tilde{P}_Q(y) = \frac{\exp\left(\mathbb{E}_{\Theta\sim Q} \log f_\theta(y)\right)}{\int \exp\left(\mathbb{E}_{\Theta\sim Q} \log f_\Theta(y)\right) d\alpha(y)} \tag{34}$$

Theorem A.2 (The reverse compensation identity)**.** *Let $\{f_\theta : \theta \in \mathcal{T}\}$ be a family of probability densities with respect to a σ-finite measure α, and suppose that $F : (\theta, y) \to f_\theta(y)$ is product measurable. Let $\Theta \sim Q$ be a $\mathcal{T}$-valued random element. Suppose that*

$$\int \exp\left(\mathbb{E}_X \log f_\theta(y)\right) d\alpha(y) > 0 \tag{35}$$

Then, for any probability measure P defined with respect to α,

$$\mathbb{E}_{\Theta\sim Q} D(P \| f_\Theta) = D(P \| \tilde{P}_Q) + \mathbb{E}_{\Theta\sim Q} D(\tilde{P}_Q \| f_\Theta) \tag{36}$$

where $\tilde{P}_Q$ is a the Q geometric mixture over $\{f_\theta\}$ with $\Theta \sim Q$.

B Proofs

Theorem 3.1 Let $\theta \sim \pi$ and $\omega \sim \pi'$ and let f_θ be a distribution with the same support as a random probability distribution P_ω for all θ and ω. Then, if $Y \sim \bar{P}$,

$$\mathop{\mathbb{E}}_{\omega\sim\pi'} \mathop{\mathbb{E}}_{\theta\sim\pi} D(P_\omega \| f_\theta) = \underbrace{D(\bar{P} \| \tilde{P})}_{\text{“squared bias”}} + \underbrace{\mathop{\mathbb{E}}_{\theta\sim\pi} D(\tilde{P} \| f_\theta)}_{\text{“variance”}} + \underbrace{I(Y;\omega)}_{\text{“irreducible error”}} \tag{37}$$

where $\bar{P}$ is the marginal

$$\bar{P}(\cdot) = \mathop{\mathbb{E}}_{\omega\sim\pi'} P_\omega(\cdot) \tag{38}$$

and where

$$\tilde{P}(\cdot) \propto \exp\left(\mathop{\mathbb{E}}_{\theta\sim\pi} \log f_\theta(\cdot)\right) \tag{39}$$

is the “geometric mixture” of f_θ over $\theta \sim \pi$.

Proof. The proof relies on the compensation identity and reverse compensation identity from information theory; see appendix A. We leverage these two identities to decompose the expected KL-divergence. Indeed, the reverse compensation identity states that for any fixed ω

$$\mathop{\mathbb{E}}_{\theta\sim\pi} D(P_\omega \| f_\theta) = D(P_\omega \| \tilde{P}) + \mathop{\mathbb{E}}_{\theta\sim\pi} D(\tilde{P} \| f_\theta) \tag{40}$$

Now we allow again for uncertainty in ω and decompose $\mathbb{E}_{\omega\sim\pi'} D(P_\omega \| \tilde{P})$. Applying the compensation identity to the right-hand side we obtain,

$$\mathop{\mathbb{E}}_{\omega\sim\pi'} \mathop{\mathbb{E}}_{\theta\sim\pi} D(P_\omega \| f_\theta) = D(\bar{P} \| \tilde{P}) + \mathop{\mathbb{E}}_{\omega\sim\pi'} D(P_\omega \| \bar{P}) + \mathop{\mathbb{E}}_{\theta\sim\pi} D(\tilde{P} \| f_\theta) \tag{41}$$

Finally we recognize $I(Y;\omega) = \mathop{\mathbb{E}}_{\omega\sim\pi'} D(P_\omega \| \bar{P})$. □

Theorem 3.3 Consider the KL-divergence loss $L(\theta, P(\Lambda)) = D(P(\Lambda)\|f_\theta)$ where Λ is data and $P(\Lambda)$ is an estimator of f_θ. Then the geometric mixture $\tilde{P}(\cdot) \propto \exp\left(\mathop{\mathbb{E}}_{\theta|\Lambda} \log f_\theta(\cdot)\right)$ is the Bayes rule; this means that $\tilde{P}(\cdot)$ is the estimator that minimizes the posterior risk:

$$\tilde{P} = \arg\min_P \mathop{\mathbb{E}}_{\theta|\Lambda} D(P\|f_\theta) \tag{42}$$

Moreover, the geometric mixture is an admissible decision rule under this loss.

Proof. Let f_θ be a probability mass function on a discrete space $\mathcal{Y}$ where $|\mathcal{Y}| = k$. Consider estimating f_θ with a function $P : \mathcal{X} \to \mathbb{S}_k$, where $\mathbb{S}_k$ is the space of k-dimensional simplexes and $X \in \mathcal{X}$ is data generated by a distribution D_θ. Suppose loss in this estimation problem is measured by the KL-divergence such that the average risk is

$$R(\theta, P) = \mathop{\mathbb{E}}_{X\sim D_\theta} D(P(X)\|f_\theta) \tag{43}$$

Given a prior $\theta \sim \pi$ we can compute the Bayes risk by integrating over $R(\theta, X)$ as follows. I exchange order of integration and denote the posterior $\theta|X \sim \pi_X$

$$r_\pi(P) = \mathop{\mathbb{E}}_{\theta\sim\pi} \mathop{\mathbb{E}}_{X\sim D_\theta} D(P(X)\|f_\theta) \tag{44}$$

$$= \mathop{\mathbb{E}}_{X\sim D} \mathop{\mathbb{E}}_{\theta\sim\pi_X} D(P(X)\|f_\theta) \tag{45}$$

Now apply the bias-variance decomposition.

$$r_\pi(P) = \mathop{\mathbb{E}}_{X\sim D}\left[D(P(X)\|\tilde{P}) + \mathop{\mathbb{E}}_{\theta\sim\pi_X} D(\tilde{P}\|f_\theta)\right] \tag{46}$$

where

$$\tilde{P}(\cdot) \propto \exp\left(\mathop{\mathbb{E}}_{\theta\sim\pi_X} \log f_\theta(\cdot)\right) \tag{47}$$

The second term in the decomposition is not a function of P; therefore, it cannot be reduced by varying the choice of P. The first term is non-negative and can be eliminated by the choice $P = \tilde{P}$. Hence, $P = \tilde{P}$ achieves the minimum Bayes average risk and is consequently the Bayes rule.

The geometric mixture is also the unique estimator with this property. To see this, observe that another estimator $\hat{P}$ would exhibit the same decomposition of the posterior risk into a reducible bias and irreducible variance. In order to minimize bias, then, the choice $\hat{P} = \tilde{P}$ (a.s.) is necessary. Finally, the geometric mixture $\tilde{P}$ is an admissible estimator because it is unique and Bayes. □

B.1 Results for Expontial Families

Results regarding the characterization of variance and the characterization of the geometric mixture using natural parameters derive from [3].

Lemma B.1 (Exponential Family Geometric Mixture)**.** *Let f_θ be in exponential family form with natural parameter $\eta = \eta(\theta)$. Then,*

$$f_\theta(\cdot) = \exp\left(\eta^\top \psi(\cdot) - A(\eta)\right) \tag{48}$$

and

$$\tilde{P}(\cdot) \propto \exp\left(\mathop{\mathbb{E}}_{\theta\sim\pi} \log f_\theta(\cdot)\right) \tag{49}$$

$$= \exp\left(\mathop{\mathbb{E}}_{\theta\sim\pi}\left[\eta^\top \psi(\cdot) - A(\eta)\right]\right) \tag{50}$$

$$\propto \exp\left(\left(\mathop{\mathbb{E}}_{\theta\sim\pi} \eta\right)^\top \psi(\cdot)\right) \tag{51}$$

from which it can be seen that $\tilde{P}$ is a member of the same exponential family with natural parameters given by the expectation under $\theta \sim \pi$.

Proposition B.2 (Characterizing Variance)**.** *When f_θ is an exponential family distribution,*

$$\mathop{\mathbb{E}}_{\theta\sim\pi} D(\tilde{P}\|f_\theta) = -A\left(\mathop{\mathbb{E}}_{\theta\sim\pi}\eta(\theta)\right) \tag{52}$$

Proof. By the previous lemma, $\tilde{P}$ is also an exponential family distribution. Hence,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\theta\sim\pi} D(\tilde{P}\|f_\theta) &= \mathop{\mathbb{E}}_{\theta\sim\pi}\left[\mathop{\mathbb{E}}_{Y\sim\tilde{P}}[-\log f_\theta(Y)] - H(\tilde{P})\right] && (53)\\
&= -\mathop{\mathbb{E}}_{Y\sim\tilde{P}}\mathop{\mathbb{E}}_{\theta\sim\pi}\log f_\theta(Y) - H(\tilde{P}) && (54)\\
&= -\mathop{\mathbb{E}}_{Y\sim\tilde{P}}\log\tilde{P}(Y) - A\left(\mathop{\mathbb{E}}_{\theta\sim\pi}\eta(\theta)\right) - H(\tilde{P}) && (55)\\
&= -A\left(\mathop{\mathbb{E}}_{\theta\sim\pi}\eta(\theta)\right) && (56)
\end{aligned}$$

□

Proposition B.3 (Characterizing Bias)**.** *When P is a distribution with zero entropy (e.g. a one-hot distribution),*

$$D(P\|\tilde{P}) = -\log\tilde{P}(y_P) \tag{57}$$

where y_P is the almost-sure outcome under P. This is the cross entropy loss attained by $\tilde{P}$.

Proof. Because P has zero entropy and hence

$$\begin{aligned}
D(P\|\tilde{P}) &= H(P,\tilde{P}) - H(P) && (58)\\
&= H(P,\tilde{P}) && (59)\\
&= -\mathop{\mathbb{E}}_{Y\sim P}\log\tilde{P}(Y) && (60)\\
&= -\log\tilde{P}(y_P) && (61)\\
&= -\mathop{\mathbb{E}}_{\theta\sim\pi}\log f_\theta(y_P) + A\left(\mathop{\mathbb{E}}_{\theta\sim\pi}\eta(\theta)\right) && (62)\\
&= \mathop{\mathbb{E}}_{\theta\sim\pi} D(P\|f_\theta) + A\left(\mathop{\mathbb{E}}_{\theta\sim\pi}\eta(\theta)\right) && (63)
\end{aligned}$$

□