On Pitfalls in OoD Detection: Predictive Entropy Considered Harmful

Andreas Kirsch [1] **Jishnu Mukhoti** [1 2] **Joost van Amersfoort** [1] **Philip H.S. Torr** [2] **Yarin Gal** [1]

Abstract

Entropy of a predictive distribution averaged over an ensemble or several posterior weight samples is often used as a metric for Out-of-Distribution (OoD) detection. However, we show that predictive entropy is inappropriate for this task because it mistakes ambiguous in-distribution samples as OoD. This issue remains hidden on curated datasets commonly used for benchmarking. We introduce a new dataset, Dirty-MNIST, with a long tail of ambiguous samples, which exemplifies this problem. Additionally, we look at the entropy of single, deterministic, softmax models and show that it is unreliable *exactly* for OoD samples. In summary, we caution against using predictive or softmax entropy for OoD detection in practice and introduce several methods to evaluate the quantitative difference between several uncertainty metrics.

1. Introduction

For the reliable detection of out-of-distribution (OoD) samples, we need to avoid mistaking ambiguous in-distribution (iD) samples as OoD. This is particularly important for noisy and ambiguous datasets found in safety-critical applications like autonomous driving (Huang & Chen, 2020) and medical diagnosis (Esteva et al., 2017; Filos et al., 2019).

OoD samples have high epistemic uncertainty, and ambiguous iD samples have high aleatoric uncertainty: *epistemic uncertainty* is inherent to the model, caused by a lack of relevant training data, and hence reducible with more data[1]; while *aleatoric uncertainty* is caused by inherent noise or ambiguity in the data and hence irreducible with more data (Der Kiureghian & Ditlevsen, 2009; Kendall & Gal, 2017). Hence, we must distinguish between the epistemic and aleatoric uncertainty of a model for a given sample, and we will see that predictive entropy confounds the two.

In deep learning, Deep Ensembles (Lakshminarayanan et al., 2017) are often used for OoD detection as they have been shown to perform well on a range of uncertainty prediction tasks. Deep Ensembles consist of several deep networks, trained independently with different random seeds, whose predictions are averaged at test time. While deep ensembling provides a small accuracy increase, it often improves uncertainty estimation as measured in OoD detection and robustness to distribution shift (Ovadia et al., 2019). A frequently used benchmark is to distinguish between two data sets based on sample uncertainty: the iD data set that was used for training, and an OoD set that is significantly different but non-trivial to distinguish. In Deep Ensembles (and related methods), the entropy of their predictive distribution is used for this benchmark (Hendrycks & Gimpel, 2016; Vyas et al., 2018; Linmans et al., 2020; Macêdo et al., 2019; Chan et al., 2020; Vernekar et al., 2019).

In this paper, we argue that firstly, using entropy for OoD detection is inherently inappropriate because it cannot distinguish between aleatoric uncertainty of ambiguous iD samples and the epistemic uncertainty of OoD samples, and secondly that the softmax entropy of a single model is even more inappropriate because it is unreliable *specifically* for samples with high epistemic uncertainty, i.e., OoD samples. This is tied to the very reason why a Deep Ensembles' mutual information captures epistemic uncertainty well and can be used to detect adversarial examples and OoD data, too (Smith & Gal, 2018). To exemplify the issues, we introduce Dirty-MNIST as a dataset with a long tail of ambiguous samples, and we analyze the relationship between softmax entropy and predictive entropy in more detail.

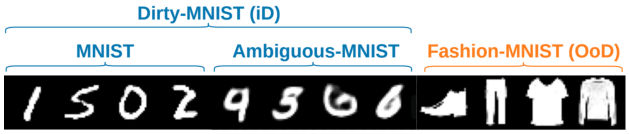
To show that entropy is inappropriate for OoD detection, we train a LeNet (LeCun et al., 1998), a VGG-16 (Simonyan & Zisserman, 2015) and a ResNet-18 with spectral normalisation, ResNet+SN[2](He et al., 2016; Miyato et al., 2018) on *Dirty-MNIST*, a modified version of MNIST (LeCun

[1]OATML, University of Oxford [2]Torr Vision Group, University of Oxford. Correspondence to: Andreas Kirsch <andreas.kirsch@cs.ox.ac.uk>.

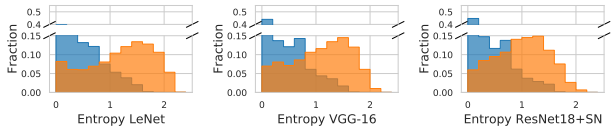
Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning., Copyright 2021 by the author(s).

[1]We follow the definition of epistemic uncertainty at input x as a quantity which is high for a previously unseen x, and decreases when x's label is added to the training set and the model is updated.

[2]Liu et al. (2020) show that spectral normalization regularizes the latent space in a way that is beneficial for OoD detection, so we also include a model trained on this recent approach.



(a) Dirty-MNIST (iD) and Fashion-MNIST (OoD)



(b) Softmax entropy

Figure 1: **(a)**: *Dirty-MNIST* consists of MNIST (60k samples) and a long tail of ambiguous samples (60k samples). **(b):** Neural networks (here *LeNet*, *VGG* & *ResNet+SN*), trained on *Dirty-MNIST as in-distribution (iD) dataset with ambiguous samples*, yield arbitrary softmax entropies for *Fashion-MNIST as OoD dataset*. OoD samples cannot be separated well from iD samples.

et al., 1998) with additional ambiguous digits (Ambiguous-MNIST), depicted in Figure 1(a). We refer to Appendix A for details on how this dataset was generated. Dirty-MNIST poses a challenge for using entropy for OoD detection as it confounds aleatoric and epistemic uncertainty: Figure 1(b) shows that the softmax entropy of a deterministic model is unable to distinguish between iD (Dirty-MNIST) and OoD (Fashion-MNIST (Xiao et al., 2017)) samples as the entropy for the latter heavily overlaps with the entropy for Ambiguous-MNIST samples. With ambiguous data having various levels of aleatoric uncertainty, Dirty-MNIST is more representative of real-world datasets compared to well-cleaned curated datasets, like MNIST and CIFAR-10, commonly used for benchmarking (Krizhevsky et al., 2009).

2. Background

In this section, we introduce Bayesian models within deep learning and relate them to Deep Ensembles and deterministic models.

Bayesian Models (Neal, 2012; MacKay, 1992) provide a principled way of measuring uncertainty. Starting with a prior distribution $p(\omega)$ over model parameters ω, they infer a posterior $p(\omega|\mathcal{D})$, given the training data $\mathcal{D}$. The predictive distribution $p(y|x,\mathcal{D})$ for a given input x is computed via marginalisation over the posterior: $p(y|x,\mathcal{D}) = \mathbb{E}_{\omega\sim p(\omega|\mathcal{D})}[p(y|x,\omega)]$. As mentioned in Gal (2016) and Smith & Gal (2018), the predictive entropy $\mathbb{H}[Y|x,\mathcal{D}]$ of $p(y|x,\mathcal{D})$ upper-bounds the epistemic uncertainty, where epistemic uncertainty is quantified as the mutual information $\mathbb{I}[Y;\omega|x,\mathcal{D}]$ (*expected information gain*) between parameters ω and output y, following the equation:

$$\underbrace{\mathbb{H}[Y|x,\mathcal{D}]}_{\text{predictive}} = \underbrace{\mathbb{I}[Y;\omega|x,\mathcal{D}]}_{\text{epistemic}} + \underbrace{\mathbb{E}_{p(\omega|\mathcal{D})}[\mathbb{H}[Y|x,\omega]]}_{\text{aleatoric (for iD } x)}. \quad (1)$$

Predictive entropy will be high for both iD ambiguous samples (high aleatoric uncertainty) and for OoD samples (high epistemic uncertainty). Hence, predictive entropy is a good measure for OoD detection only when used with curated datasets that do not contain ambiguous samples, unlike Dirty-MNIST in Figure 1. Note that aleatoric uncertainty is only meaningful in-distribution because it quantifies the level of ambiguity between the different classes which might be observed for input x. If the probability of observing x under the data generating distribution is zero, the probability $p(y|x) = \frac{p(x,y)}{p(x)}$ is not defined, and its entropy as a measure of aleatoric uncertainty is not defined.

Deep Ensembles (Lakshminarayanan et al., 2017) are an ensemble of neural networks which average the models' softmax outputs. Uncertainty is then estimated as the entropy of this averaged softmax vector. Note that ensembling might also be seen as performing Bayesian Model Averaging (He et al., 2020; Wilson & Izmailov, 2020), as each ensemble member, producing a softmax output $p(y|x,\omega)$, can be considered to be drawn from some distribution $p(\omega|\mathcal{D})$ over the trained model parameters ω, which is induced by the pushforward of the weight initialization under stochastic optimization. As a result, Equation (1) can also be applied to Deep Ensembles to disentangle epistemic from predictive uncertainty.

In practice, both mutual information $\mathbb{I}[Y;\omega|x,\mathcal{D}]$ and predictive entropy $\mathbb{H}[Y|x,\mathcal{D}]$ are used in the literature to detect OoD samples, but predictive entropy will be high whenever either epistemic uncertainty is high, or when aleatoric uncertainty is high: it upper bounds the mutual information. This can help separate iD and OoD data better for curated iD datasets, offering an explanation for previous empirical findings of predictive entropy outperforming mutual information (Malinin & Gales, 2018). With ambiguous iD samples, it can lead to confounding, however.

Deterministic Models produce a softmax distribution $p(y|x,\omega)$ and use either the maximum softmax probability $\max_c p(y=c|x,\omega)$ (*confidence*) or the *softmax entropy* $\mathbb{H}[Y|x,\omega]$ as a measure of uncertainty. It is well-known that these measures are often not indicative of OoD data (Hendrycks & Gimpel, 2016; Guo et al., 2017). Popular approaches to tackle this problem include pre-processing of inputs and post-hoc calibration methods (Liang et al., 2018; Guo et al., 2017), alternative objective functions (Lee et al., 2018; DeVries & Taylor, 2018), and exposure to outliers (Hendrycks et al., 2018). However, these methods suffer from several shortcomings including failing to perform under distribution shift (Ovadia et al., 2019), requiring significant changes to the training setup, and assuming the availability of OoD samples during training (which many applications do not have access to). In the next section, we demonstrate that the softmax entropy is inherently inappropriate to capture epistemic uncertainty.

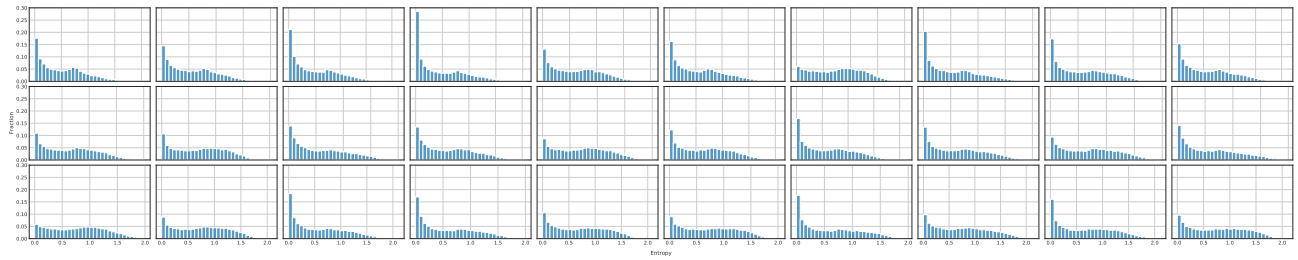


Figure 2: *Softmax entropy histograms of 30* Wide-ResNet-28-10+SN *models trained on CIFAR-10, evaluated on SVHN (OoD).* The softmax entropy distribution of the different models varies considerably.

3. Qualitative & Quantitative Statements

In this section, we show that *softmax entropy cannot capture epistemic uncertainty exactly because a Deep Ensemble's mutual information can*. Proofs are provided in Appendix D.

As mentioned in Section 2, Equation (1) can be used with Deep Ensembles, as each ensemble member can be considered a sample from *some* distribution $\mathrm{p}(\omega \mid \mathcal{D})$ over model parameters $\omega \subset \Omega$ (e.g. a uniform distribution over K trained ensemble members $\omega_1, ..., \omega_K$). Note that the mutual information $\mathbb{I}[Y; \omega \mid x, \mathcal{D}]$ isolates epistemic for Deep Ensembles as well, whereas the predictive entropy $\mathbb{H}[Y \mid x, \mathcal{D}]$ (often used with Deep Ensembles) measures predictive uncertainty, which will be high whenever either epistemic or aleatoric uncertainties are high[3]. Furthermore, the mechanism underlying Deep Ensemble uncertainty that pushes epistemic uncertainty to be high on OoD data is the function disagreement between different ensemble components, i.e. arbitrary extrapolations of the softmax models composing the ensemble (leading the "aleatoric" term in Equation (1) to vanish (Smith & Gal, 2018)). We can formalize this intuition in the following quantitative statement using Equation (1):

Proposition 3.1. *Let x_1 and x_2 be points such that x_1 has **higher** epistemic uncertainty than x_2 under the ensemble: $\mathbb{I}[Y_1; \omega \mid x_1, \mathcal{D}] > \mathbb{I}[Y_2; \omega \mid x_2, \mathcal{D}] + \delta$, with $\delta \geq 0$. Further assume both have similar predictive entropy $|\mathbb{H}[Y_1 | x_1, \mathcal{D}] - \mathbb{H}[Y_2 \mid x_2, \mathcal{D}]| \leq \epsilon$, for $\epsilon \geq 0$. Then, there exist sets of ensemble members $\hat{\Omega}$ with $\mathrm{p}(\hat{\Omega} \mid \mathcal{D}) > 0$, such that for all softmax models $\omega \in \hat{\Omega}$ the softmax entropy of x_1 is **lower** than the softmax entropy of x_2: $\mathbb{H}[Y_1 \mid x_1, \omega] < \mathbb{H}[Y_2 \mid x_2, \omega] - (\delta - \epsilon)$.*

This shows that if a sample is assigned higher epistemic uncertainty (in the form of mutual information) by a Deep Ensemble, it will necessarily be assigned lower softmax entropy by at least one of the ensemble's members. As a result, *the empirical observation that the mutual information of an ensemble can quantify epistemic uncertainty well implies that the softmax entropy of a deterministic model cannot*. This claim is further supported by Figure 1(b) (and Appendix C and D.3) where we observe the softmax entropy for OoD samples to have high variance.

[3] A mutual information estimator based on a Deep Ensemble's outputs will lower-bound the actual mutual information. This explains why the predictive entropy performs better: its estimate upper-bounds the estimated mutual information and is thus likely to be closer to the actual mutual information/epistemic uncertainty.

4. Quantitative Modeling & Empirical Validation

To gain further insights, we analyze the relationship between softmax entropies and predictive entropies and quantify the variance of softmax entropies given the predictive distribution and epistemic uncertainty in the ensemble setting. We empirically find that a Dirichlet distribution lower-bounds the variance, but models the entropy distribution well. This preliminary result allows us quantify how strongly the softmax entropies vary across different deterministic models.

For a distribution over models $\mathrm{p}(\omega \mid \mathcal{D})$, and a sample x, we have $\mathrm{p}(y \mid x)$, $\mathbb{I}[Y; \omega \mid x]$. We use moment matching with these two quantities to fit a Dirichlet distribution $\mathbf{p} \sim \mathrm{Dir}(\alpha(x, \omega))$ as a model for $\mathrm{p}(y \mid x, \omega)$, which satisfies:

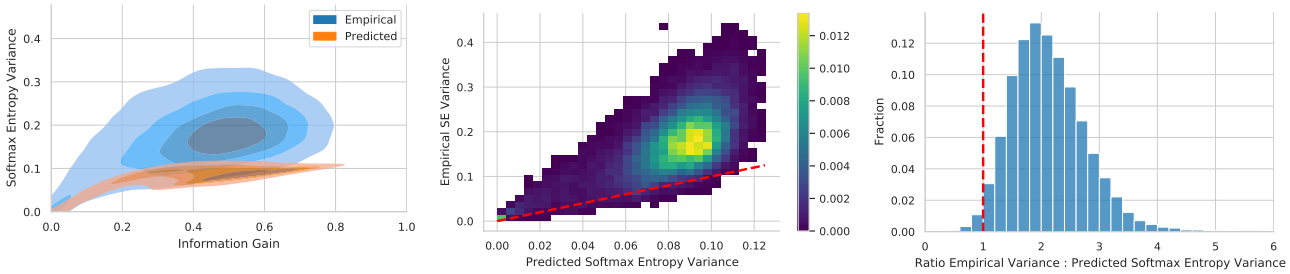
$$\mathrm{p}(y \mid x) = \frac{\alpha_i}{\alpha_0} \tag{2}$$

$$\mathbb{H}[Y \mid x] - \mathbb{I}[Y; \omega \mid x] = \psi(\alpha_0 + 1) - \sum_{y=1}^{K} \mathrm{p}(y \mid x) \psi(\alpha_0\ \mathrm{p}(y \mid x) + 1). \tag{3}$$

We can thus compute the variance $\mathrm{Var}[\mathbb{H}[Y \mid \mathbf{p}]]$ (see appendix) as a proxy for $\mathrm{Var}[\mathbb{H}[Y \mid x, \omega]]$.

Empirical results show that $\mathrm{Var}[\mathbb{H}[Y \mid \mathbf{p}]]$ lower-bounds $\mathrm{Var}[\mathbb{H}[Y \mid x, \omega]]$ for common models in Figure 3. This simple approximation seems to be able to capture the empirical entropy distribution quite well as shown in Figure 4. In particular, we see in Figure 3(a) that most OoD samples have epistemic uncertainty around 0.5 nats with considerable variance of around ≈ 0.15 nats^2 (std dev ≈ 0.4 nats).

We train an ensemble of VGG and WideResNet-28-10+SN models (25 members each) on CIFAR-10 and compute the predictive entropy, mutual information, and softmax entropies for each sample in SVHN (as OoD distribution with

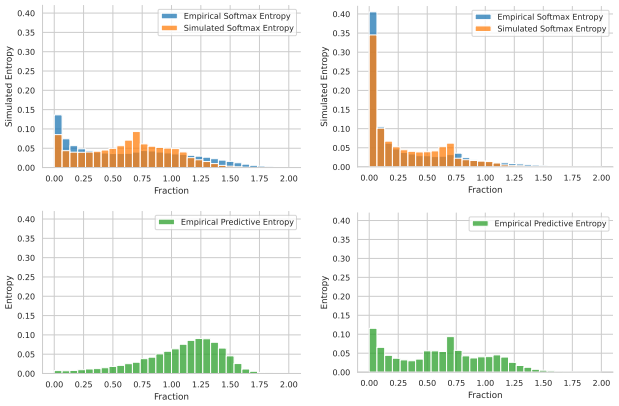


(a) Epistemic Uncertainty vs Softmax Entropy Variance

(b) Verification

(c) Ratio

Figure 3: *The variance of softmax entropies can be lower-bounded by fitting Dirichlet distributions on the OoD samples* $\mathrm{p}(y \mid x, \omega)$ *from SVHN for a WideResNet-28-10+SN trained on CIFAR-10.* **(a)** The variance of the softmax entropy is not linearly correlated to the epistemic uncertainty. For both high and low epistemic uncertainty, the variance decreases. **(b)** The empirical variance of softmax entropies is effectively lower-bounded by the predicted variance of the Dirichlet distributions. The red dashed line depicts equality. **(c)** The ratio histogram shows that there are only few violations due to precision issues.



(a) WideResNet-28-10+SN

(b) VGG16

Figure 4: *Simulated vs empirical softmax entropy and (ensemble) predictive entropy on WideResNet-28-10+SN and VGG16.* Although we use a simple Dirichlet model, sampling from the fitted Dirichlet distributions does approximate the empirical entropy distribution well.

ambiguous data). Figure 2 depicts how the softmax entropies vary considerably across the different ensemble members over SVHN. In Figure 3, we see empirically that the softmax entropy variance of the fitted Dirichlet distributions generally lower-bounds the true softmax entropy variance. Morever, Figure 3(a) shows both **i)** the non-linear relationship between epistemic uncertainty and variance in the softmax entropies and **ii)** that Dirichlet distributions cannot capture this well. However, the fitted Dirichlet distributions provide a good proxy for the overall distribution of softmax entropies: we draw softmax distributions from the Dirichlets for each sample and compare the respective entropies to the true softmax entropies on SVHN in Figure 4. For both VGG and WideResNet, we see that the histograms are close, so even though this quantitive model is very simple, it captures the true distribution of softmax entropies surprisingly well.

5. Conclusion & Limitations

Both through quantitative and qualitative statements as well as through empirical validation, we show that neither the predictive entropy of Deep Ensembles nor the softmax entropy of deterministic models is appropriate for measuring epistemic uncertainty and OoD detection tasks. This holds in particularly for real world datasets that contain more ambiguous data than the curated datasets that are employed for benchmarking. Proposition 3.1 shows that it does not provide a stable ranking of points by epistemic uncertainty, and that indeed, the predictive entropy of Deep Ensembles captures epistemic uncertainty through the variance of the softmax entropies of deterministic models. Additionally, we have examined a simple model that relates the predictive entropy and epistemic uncertainty (through the mutual information) of Deep Ensembles with the softmax entropies of deterministic models.

There are two limitations: first, the Dirichlet distribution that we fit is not a maximum entropy distribution given the constraints we pose, as far as we know. It is surprising that it provides a good lower bound empirically. At the same time, this simple model seems to match the empirical softmax entropy distribution on OoD data well. Further work is needed to establish and prove true lower bounds or approximations. Second, while our argument is true for epistemic uncertainty generally, it does not generalize to OoD detection tasks when the model is trained with "OoD" hold-out data of any form as this breaks the implied equivalence "*OoD data* $\iff$ *high epistemic uncertainty*". Indeed, training using OoD data shifts epistemic uncertainty into aleatoric uncertainty and can be captured by the softmax entropy of deterministic models as well as the predictive entropy of a Deep Ensemble—even though it confounds ambiguous iD samples with OoD samples. However, is such data still truly OoD when we start training or fine-tuning on it, or are we simply moving the goal posts?

REFERENCES

Chan, R., Rottmann, M., and Gottschalk, H. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. *arXiv preprint arXiv:2012.06575*, 2020.

Der Kiureghian, A. and Ditlevsen, O. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

DeVries, T. and Taylor, G. W. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Filos, A., Farquhar, S., Gomez, A. N., Rudner, T. G., Kenton, Z., Smith, L., Alizadeh, M., de Kroon, A., and Gal, Y. A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks. *arXiv preprint arXiv:1912.10481*, 2019.

Gal, Y. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*, 2017.

He, B., Lakshminarayanan, B., and Teh, Y. W. Bayesian Deep Ensembles via the Neural Tangent Kernel. In *Advances in neural information processing systems*, 2020.

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.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.

Huang, Y. and Chen, Y. Autonomous driving with deep learning: A survey of state-of-art technologies. *arXiv preprint arXiv:2006.06091*, 2020.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pp. 5574–5584, 2017.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *ICLR*, 2014.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pp. 6402–6413, 2017.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Linmans, J., van der Laak, J., and Litjens, G. Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks. In *Medical Imaging with Deep Learning*, 2020. URL https://openreview.net/forum?id=IdZWfAGuuB.

Liu, J. Z., Lin, Z., Padhy, S., Tran, D., Bedrax-Weiss, T., and Lakshminarayanan, B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *NeurIPS*, 2020.

Macêdo, D., Ren, T. I., Zanchettin, C., Oliveira, A. L., and Ludermir, T. Isotropy maximization loss and entropic score: Accurate, fast, efficient, scalable, and turnkey neural networks out-of-distribution detection based on the principle of maximum entropy. *arXiv preprint arXiv:1908.05569*, 2019.

MacKay, D. J. *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992.

Malinin, A. and Gales, M. J. Predictive uncertainty estimation via prior networks. In *NeurIPS*, 2018.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Neal, R. M. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

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. In *Advances*

in Neural Information Processing Systems, pp. 13991–14002, 2019.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Smith, L. and Gal, Y. Understanding Measures of Uncertainty for Adversarial Example Detection. In *UAI*, 2018.

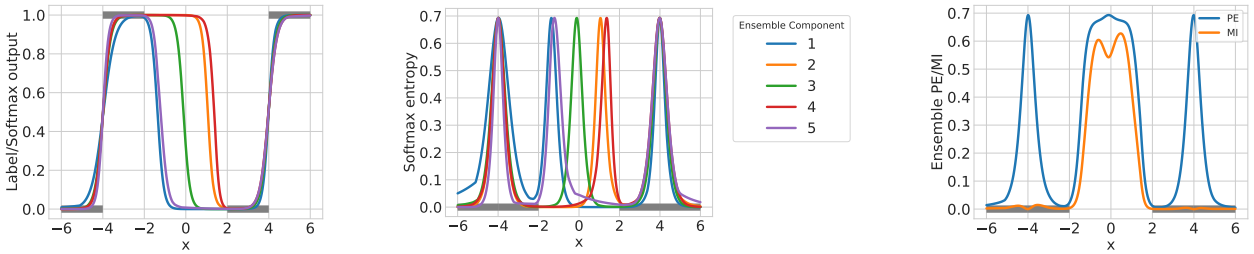
Vernekar, S., Gaurav, A., Abdelzad, V., Denouden, T., Salay, R., and Czarnecki, K. Out-of-distribution detection in classifiers via generation. *arXiv preprint arXiv:1910.04241*, 2019.

Vyas, A., Jammalamadaka, N., Zhu, X., Das, D., Kaul, B., and Willke, T. L. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 550–564, 2018.

Wilson, A. G. and Izmailov, P. Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*, 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.



(a) Softmax output (b) Softmax entropy (c) PE and MI of ensemble

Figure 5: *Visualisation of softmax outputs/entropies along with the Predictive Entropy (PE) and Mutual Information of a 5-Ensemble.* Figures (a) and (b) show that the softmax entropy is high at points of ambiguity, i.e., where the label changes from 0 to 1 for the data, thereby capturing aleatoric uncertainty, whereas softmax entropy can be low or high for OoD (between -2 to 2). At the same time, figure (c) shows that the MI of the ensemble is only high for OoD, thereby solely capturing epistemic uncertainty, whereas the PE of the ensemble is high for both OoD and for regions of ambiguity, thereby capturing both epistemic and aleatoric uncertainty.

A. Ambiguous- and Dirty-MNIST

Each sample in Ambiguous-MNIST is constructed by decoding a linear combination of latent representations of 2 different MNIST digits from a pre-trained VAE (Kingma & Welling, 2014). Every decoded image is assigned several labels sampled from the softmax probabilities of an off-the-shelf MNIST neural network ensemble, with points filtered based on an ensemble's MI (to remove 'junk' images). All off-the-shelf MNIST neural networks were then discarded and new models were trained to generate Figure 1 (and as can be seen, the ambiguous points we generate indeed have high entropy regardless of the model architecture used). We create 60K training and 10K test images for Ambiguous-MNIST. Finally, the Dirty-MNIST dataset in this experiment contains MNIST and Ambiguous-MNIST samples in a 1:1 ratio (with 120K training and 20K test samples).

B. Experimental Details

We use Wide-ResNet-28-10 (Zagoruyko & Komodakis, 2016) as the model architecture for all the baselines. We train the softmax baselines for 350 epochs using SGD as the optimiser with a momentum of 0.9, and an initial learning rate of 0.1. The learning rate drops by a factor of 10 at epochs 150 and 250. We train the 5-Ensemble baseline using this same training setup.

C. 5-Ensemble Visualisation

In Figure 5, we provide a visualisation of a 5-ensemble (with five deterministic softmax networks) to see how softmax entropy fails to capture epistemic uncertainty precisely because the mutual information (MI) of an ensemble does not. We train the networks on 1-dimensional data with binary labels 0 and 1. The data is shown in Figure 5(b). From Figure 5(a) and Figure 5(b), we find that softmax entropy is high in regions of ambiguity where the label changes from 0 to 1 (i.e., at x value -4 and 4). This indicates that softmax entropy can capture aleatoric uncertainty. Furthermore, in the x interval $(-2, 2)$, we find that the deterministic softmax networks disagree in their predictions (see Figure 5(a)) and have softmax entropies which can be high, low or anywhere in between (see Figure 5(b)). In fact, this disagreement is the very reason why the MI of the ensemble is high in the interval $(-2, 2)$, thereby reliably capturing epistemic uncertainty. Finally, note that the predictive entropy (PE) of the ensemble is high both in the OoD interval $(-2, 2)$ as well as at points of ambiguity (i.e., at -4 and 4). This indicates that the PE of a Deep Ensemble captures both epistemic and aleatoric uncertainty well. From these visualisations, we draw the conclusion that the softmax entropy of a deterministic softmax model cannot capture epistemic uncertainty precisely because the MI of a Deep Ensemble can.

D. Theoretical Results

D.1. Qualitative Statement

We start with a proof of Proposition 3.1, which quantitatively examines the qualitative statemets that given the same predictive entropy, higher epistemic uncertainty for one point than another will cause some ensemble members to have lower softmax entropy.

Proposition 3.1. *Let x_1 and x_2 be points such that x_1 has **higher** epistemic uncertainty than x_2 under the ensemble: $\mathbb{I}[Y_1; \omega \mid x_1, \mathcal{D}] > \mathbb{I}[Y_2; \omega \mid x_2, \mathcal{D}] + \delta$, with $\delta \geq 0$. Further assume both have similar predictive entropy $|\, \mathbb{H}[Y_1 \mid x_1, \mathcal{D}] - \mathbb{H}[Y_2 \mid x_2, \mathcal{D}]| \leq \epsilon$, for $\epsilon \geq 0$. Then, there exist sets of ensemble members $\hat{\Omega}$ with $\mathrm{p}(\hat{\Omega} \mid \mathcal{D}) > 0$, such that for all softmax models $\omega \in \hat{\Omega}$ the softmax entropy of x_1 is **lower** than the softmax entropy of x_2: $\mathbb{H}[Y_1 \mid x_1, \omega] < \mathbb{H}[Y_2 \mid x_2, \omega] - (\delta - \epsilon)$.*

Proof. From Equation (1), we obtain

$$\begin{aligned} &|\, \mathbb{I}[Y_1; \omega \mid x_1, \mathcal{D}] + \mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D})} \left[\mathbb{H}[Y_1 \mid x_1, \omega]\right] \\ &- \mathbb{I}[Y_2; \omega \mid x_2, \mathcal{D}] - \mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D})} \left[\mathbb{H}[Y_2 \mid x_2, \omega]\right] \,| \leq \epsilon. \end{aligned} \tag{4}$$

and hence we have

$$\begin{aligned} &\mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D})} \left[\mathbb{H}[Y_1 \mid x_1, \omega]\right] - \mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D})} \left[\mathbb{H}[Y_2 \mid x_2, \omega]\right] \\ &+ \underbrace{\left(\mathbb{I}[Y_1; \omega \mid x_1, \mathcal{D}] - \mathbb{I}[Y_2; \omega \mid x_2, \mathcal{D}]\right)}_{> \delta} \leq \epsilon. \end{aligned} \tag{5}$$

We rearrange the terms:

$$\mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D})} \left[\mathbb{H}[Y_1 \mid x_1, \omega]\right] < \mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D})} \left[\mathbb{H}[Y_2 \mid x_2, \omega]\right] - (\delta - \epsilon). \tag{6}$$

Now, the statement follows by contraposition: if $\mathbb{H}[Y_1 \mid x_1, \omega] \geq \mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D})} \left[\mathbb{H}[Y_2 \mid x_2, \omega]\right] - (\delta - \epsilon)$ for all ω, the monotonicity of the expectation would yield $\mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D})} \left[\mathbb{H}[Y_1 \mid x_1, \omega]\right] \geq \mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D})} \left[\mathbb{H}[Y_2 \mid x_2, \omega]\right] - (\delta - \epsilon)$. Thus, there is a non-null-set Ω' with $\mathrm{p}(\Omega') > 0$, such that

$$\mathbb{H}[Y_1 \mid x_1, \omega] < \mathbb{H}[Y_2 \mid x_2, \omega] - (\delta - \epsilon), \tag{7}$$

for all $\omega \in \Omega'$. □

While this statement provides us with an intuition for why ensemble members and thus deterministic models cannot provide epistemic uncertainty reliably through their softmax entropies, we can examine this further by establishing some upper bounds.

D.2. Infinite Deep Ensemble

There are two interpretations of the ensemble parameter distribution $\mathrm{p}(\omega \mid \mathcal{D})$: we can view it as an empirical distribution given a specific ensemble with members $\omega_{i \in \{1, \ldots, K\}}$, or we can view it as a distribution over all possible trained models, given: random weight initializations, the dataset, stochasticity in the minibatches and the optimization process. In that case, any Deep Ensemble with K members can be seen as finite Monte-Carlo sample of this posterior distribution. The predictions of an ensemble then are an unbiased estimate of the predictive distribution $\mathbb{E}_{\mathrm{p}(\omega \mid \mathcal{D})} \left[\mathrm{p}(y \mid x, \omega|)\right]$, and similarly the expected information gain computed using the members of the Deep Ensemble is just a (biased) estimator of $\mathbb{I}[Y; \omega \mid x, \mathcal{D}]$.

D.3. Analysis of Softmax Entropy of a Single Deterministic Model on OoD Data using Properties of Deep Ensembles

Based on the interpretation of Deep Ensembles as a distribution over model parameters, we can walk backwards and, given *some value* for the predictive distribution and epistemic uncertainty of a Deep Ensemble, estimate what the softmax entropies from each ensemble component must have been. I.e. if we observe Deep Ensembles to have high epistemic uncertainty on OoD data, we can deduce from that what the softmax entropy of deterministic neural nets (the ensemble components) must look like. More specifically, given a predictive distribution $\mathrm{p}(y \mid x)$ and epistemic uncertainty, that is expected information gain $\mathbb{I}[Y; \omega \mid x]$, of the infinite Deep Ensemble, we estimate the expected softmax entropy from a single deterministic model, considered as a sample $\omega \sim \mathrm{p}(\omega \mid \mathcal{D})$ and model a lower bound for the variance. Empirically, we find the variance to be

higher by a large amount for OoD samples, showing that softmax entropies do not capture epistemic uncertainty well for samples with high epistemic uncertainty.

We will need to make several strong assumptions that limit the generality of our estimation, but we can show that our analysis models the resulting softmax entropy distributions appropriately. This will show that deterministic softmax models can have widely different entropies and confidence values.

Given the predictive distribution $\mathrm{p}(y \mid x)$ and epistemic uncertainty $\mathbb{I}[Y; \omega \mid x]$, we can approximate the distribution over softmax probability vectors $p(y|x, \omega)$ for different ω using its maximum-entropy estimate: a Dirichlet distribution $(Y_1, \ldots, Y_K) \sim \mathrm{Dir}(\alpha)$ with non-negative concentration parameters $\alpha = (\alpha_1, \ldots, \alpha_K)$ and $\alpha_0 := \sum \alpha_i$. Note that the Dirichlet distribution is used *only as an analysis tool*.

D.3.1. PRELIMINARIES

Before we can establish our main result, we need to look more closely at Dirichlet-Multinomial distributions. Given a Dirichlet distribution $\mathrm{Dir}(\alpha)$ and a random variable $\mathbf{p} \sim \mathrm{Dir}(\alpha)$, we want to quantify the expected entropy $\mathbb{E}_{\mathbf{p}\sim\mathrm{Dir}(\alpha)}\, \mathbb{H}_{Y\sim\mathrm{Cat}(\mathbf{p})}[Y]$ and its variance $\mathrm{Var}_{\mathbf{p}\sim\mathrm{Dir}(\alpha)}\, \mathbb{H}_{Y\sim\mathrm{Cat}(\mathbf{p})}[Y]$. For this, we need to develop more theory. In the following, Γ denotes the Gamma function, ψ denotes the Digamma function, ψ' denotes the Trigamma function.

Lemma D.1. *Given a Dirichlet distribution and random variable $\mathbf{p} \sim \mathrm{Dir}(\alpha)$, the following hold:*
1. The expectation $\mathbb{E}[\log \mathbf{p}_i]$ is given by:

$$\mathbb{E}[\log \mathbf{p}_i] = \psi(\alpha_i) - \psi(\alpha_0). \tag{8}$$

2. The covariance $\mathrm{Cov}[\log \mathbf{p}_i, \log \mathbf{p}_j]$ is given by

$$\mathrm{Cov}[\log \mathbf{p}_i, \log \mathbf{p}_j] = \psi'(\alpha_i)\, \delta_{ij} - \psi'(\alpha_0). \tag{9}$$

3. The expectation $\mathbb{E}\left[\mathbf{p}_i^n \mathbf{p}_j^m \log \mathbf{p}_i\right]$ is given by:

$$\begin{aligned} &\mathbb{E}\left[\mathbf{p}_i^n \mathbf{p}_j^m \log \mathbf{p}_i\right] \\ &\quad = \frac{\alpha_i^{\overline{n}}\, \alpha_j^{\overline{m}}}{\alpha_0^{\overline{n+m}}} \left(\psi(\alpha_i + n) - \psi(\alpha_0 + n + m)\right), \end{aligned} \tag{10}$$

where $i \neq j$, and $n^{\overline{k}} = n\,(n+1)\,\ldots\,(n+k-1)$ denotes the rising factorial.

Proof. 1. The Dirichlet distribution is members of the exponential family. Therefore the moments of the sufficient statistics are given by the derivatives of the partition function with respect to the natural parameters. The natural parameters of the Dirichlet distribution are just its concentration parameters α_i. The partition function is

$$A(\alpha) = \sum_{i=1}^{k} \log \Gamma(\alpha_i) - \log \Gamma(\alpha_0), \tag{11}$$

the sufficient statistics is $T(x) = \log x$, and the expectation $\mathbb{E}[T]$ is given by

$$\mathbb{E}[T_i] = \frac{\partial A(\alpha)}{\partial \alpha_i} \tag{12}$$

as the Dirichlet distribution is a member of the exponential family. Substituting the definitions and evaluating the partial derivative yields

$$\mathbb{E}[\log \mathbf{p}_i] = \frac{\partial}{\partial \alpha_i} \left[\sum_{i=1}^{k} \log \Gamma(\alpha_i) - \log \Gamma\left(\sum_{i=1}^{k} \alpha_i\right)\right] \tag{13}$$

$$= \psi(\alpha_i) - \psi(\alpha_0)\, \frac{\partial}{\partial \alpha_i} \alpha_0, \tag{14}$$

where we have used that the Digamma function ψ is the log derivative of the Gamma function $\psi(x) = \frac{d}{dx} \ln \Gamma(x)$. This proves (8) as $\frac{\partial}{\partial \alpha_i} \alpha_0 = 1$.

2. Similarly, the covariance is obtained using a second-order partial derivative:

$$\mathrm{Cov}[T_i, T_j] = \frac{\partial^2 A(\alpha)}{\partial \alpha_i \, \partial \alpha_i}. \tag{15}$$

Again, substituting yields

$$\mathrm{Cov}[\log \mathbf{p}_i, \log \mathbf{p}_j] = \frac{\partial}{\partial \alpha_j} \left[\psi\left(\alpha_i\right) - \psi\left(\alpha_0\right)\right] \tag{16}$$

$$= \psi'\left(\alpha_i\right) \delta_{ij} - \psi'\left(\alpha_0\right). \tag{17}$$

3. We will make use of a simple reparameterization to prove the statement using Equation (8). Expanding the expectation and substituting the density $\mathrm{Dir}(\mathbf{p}; \alpha)$, we obtain

$$\mathbb{E}\left[\mathbf{p}_i^n \mathbf{p}_j^m \log \mathbf{p}_i\right] = \int \mathrm{Dir}(\mathbf{p}; \alpha)\, \mathbf{p}_i^n\, \mathbf{p}_j^m\, \log \mathbf{p}_i \, d\mathbf{p} \tag{18}$$

$$= \int \frac{\Gamma\left(\alpha_0\right)}{\prod_{i=1}^K \Gamma\left(\alpha_i\right)} \prod_{k=1}^K \mathbf{p}_k^{\alpha_k - 1}\, \mathbf{p}_i^n\, \mathbf{p}_j^m\, \log \mathbf{p}_i \, d\mathbf{p} \tag{19}$$

$$= \frac{\Gamma(\alpha_i + n)\Gamma(\alpha_j + m)\Gamma(\alpha_0 + n + m)}{\Gamma(\alpha_i)\Gamma(\alpha_j)\Gamma(\alpha_0)} \int \mathrm{Dir}(\hat{\mathbf{p}}; \hat{\alpha})\, \hat{\mathbf{p}}_i^n\, \hat{\mathbf{p}}_j^m\, \log \hat{\mathbf{p}}_i \, d\hat{\mathbf{p}} \tag{20}$$

$$= \frac{\alpha_i^{\overline{n}}\, \alpha_j^{\overline{m}}}{\alpha_0^{\overline{n+m}}}\, \mathbb{E}\left[\log \hat{\mathbf{p}}_i\right], \tag{21}$$

where $\hat{\mathbf{p}} \sim \mathrm{Dir}(\hat{\alpha})$ with $\hat{\alpha} = (\alpha_0, \ldots, \alpha_i + n, \ldots, \alpha_j + m, \ldots, \alpha_K)$ and we made use of the fact that $\frac{\Gamma(z+n)}{\Gamma(z)} = z^{\overline{n}}$. Finally, we can apply Equation (8) on $\hat{\mathbf{p}} \sim \mathrm{Dir}(\hat{\alpha})$ to show

$$= \frac{\alpha_i^{\overline{n}}\, \alpha_j^{\overline{m}}}{\alpha_0^{\overline{n+m}}} \left(\psi(\alpha_i + n) - \psi(\alpha_0 + n + m)\right). \tag{22}$$

□

With this, we can already quantify the expected entropy $\mathbb{E}_{\mathbf{p} \sim \mathrm{Dir}(\alpha)}\, \mathbb{H}_{Y \sim \mathrm{Cat}(\mathbf{p})}[Y]$:

Lemma D.2. *Given a Dirichlet distribution and a random variable* $\mathbf{p} \sim Dir(\alpha)$*, the expected entropy* $\mathbb{E}_{\mathbf{p} \sim \mathrm{Dir}(\alpha)}\, \mathbb{H}_{Y \sim \mathrm{Cat}(\mathbf{p})}[Y]$ *of the categorical distribution* $Y \sim \mathrm{Cat}(\mathbf{p})$ *is given by*

$$\mathbb{E}_{\mathrm{p}(\mathbf{p}|\alpha)}\, \mathbb{H}[Y \mid \mathbf{p}] = \psi(\alpha_0 + 1) - \sum_{y=1}^K \frac{\alpha_i}{\alpha_0} \psi(\alpha_i + 1). \tag{23}$$

Proof. Applying the sum rule of expectations and Equation (10) from Lemma D.1, we can write

$$\mathbb{E}\, \mathbb{H}[Y \mid \mathbf{p}] = \mathbb{E}\left[-\sum_{i=1}^K \mathbf{p}_i \log \mathbf{p}_i\right] = -\sum_i \mathbb{E}\left[\mathbf{p}_i \log \mathbf{p}_i\right] \tag{24}$$

$$= -\sum_i \frac{\alpha_i}{\alpha_0} \left(\psi(\alpha_i + 1) - \psi(\alpha_0 + 1)\right). \tag{25}$$

The result follows after rearranging and making use of $\sum_i \frac{\alpha_i}{\alpha_0} = 1$. □

With these statements, we can answer a slightly more complex problem:

Lemma D.3. *Given a Dirichlet distribution and a random variable* $\mathbf{p} \sim \mathrm{Dir}(\alpha)$*, the covariance* $\mathrm{Cov}[\mathbf{p}_i^n \log \mathbf{p}_i, \mathbf{p}_j^m \log \mathbf{p}_j]$ *is given by*

$$\begin{aligned}
&\mathrm{Cov}[\mathbf{p}_i^n \log \mathbf{p}_i, \mathbf{p}_j^m \log \mathbf{p}_j] && (26)\\
&= \frac{\alpha_i^{\overline{n}}\, \alpha_j^{\overline{m}}}{\alpha_0^{\overline{n+m}}} \left((\psi(\alpha_i + n) - \psi(\alpha_0 + n + m))\right.\\
&\quad (\psi(\alpha_j + m) - \psi(\alpha_0 + n + m))\\
&\quad \left. -\psi'(\alpha_0 + n + m)\right) && (27)\\
&\quad + \frac{\alpha_i^{\overline{n}}\, \alpha_j^{\overline{m}}}{\alpha_0^{\overline{n}}\, \alpha_0^{\overline{m}}} (\psi(\alpha_i + n) - \psi(\alpha_0 + n))\\
&\quad (\psi(\alpha_j + m) - \psi(\alpha_0 + n)),
\end{aligned}$$

for $i \neq j$*, where* ψ *is the Digamma function and* ψ' *is the Trigamma function. Similarly, the covariance* $\mathrm{Cov}[\mathbf{p}_i^n \log \mathbf{p}_i, \mathbf{p}_i^m \log \mathbf{p}_i]$ *is given by*

$$\begin{aligned}
&\mathrm{Cov}[\mathbf{p}_i^n \log \mathbf{p}_i, \mathbf{p}_i^m \log \mathbf{p}_i] && (28)\\
&= \frac{\alpha_i^{\overline{n+m}}}{\alpha_0^{\overline{n+m}}} \left((\psi(\alpha_i + n + m) - \psi(\alpha_0 + n + m))^2\right.\\
&\quad \left. + \psi'(\alpha_i + n + m) - \psi'(\alpha_0 + n + m)\right) && (29)\\
&\quad + \frac{\alpha_i^{\overline{n}}\, \alpha_i^{\overline{m}}}{\alpha_0^{\overline{n}}\, \alpha_0^{\overline{m}}} (\psi(\alpha_i + n) - \psi(\alpha_0 + n))\\
&\qquad (\psi(\alpha_i + m) - \psi(\alpha_0 + n)).
\end{aligned}$$

Regrettably, the equations are getting large. By abuse of notation, we introduce a convenient shorthand before proving the lemma.

Definition D.4. *We will denote by*

$$\overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_i^{n,m}\right]} = \psi(\alpha_i + n) - \psi(\alpha_0 + n + m), \tag{30}$$

and use $\overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_i^{n}\right]}$ *for* $\overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_i^{n,0}\right]}$*. Likewise,*

$$\overline{\mathrm{Cov}[\log \hat{\mathbf{p}}_i^{n,m}, \log \hat{\mathbf{p}}_j^{n,m}]} = \psi'(\alpha_i + n)\delta_{ij} - \psi'(\alpha_0 + n + m). \tag{31}$$

This notation agrees with the proof of Equation (8) and (9) in Lemma D.1. With this, we can significantly simplify the previous statements:

Corollary D.4.1. *Given a Dirichlet distribution and random variable* $\mathbf{p} \sim \mathrm{Dir}(\alpha)$*,*

$$\mathbb{E}\left[\mathbf{p}_i^n \mathbf{p}_j^m \log \mathbf{p}_i\right] = \frac{\alpha_i^{\overline{n}}\, \alpha_j^{\overline{m}}}{\alpha_0^{\overline{n+m}}} \overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_i^{n,m}\right]}, \tag{32}$$

$$\begin{aligned}
&\mathrm{Cov}[\mathbf{p}_i^n \log \mathbf{p}_i, \mathbf{p}_j^m \log \mathbf{p}_j] && (33)\\
&= \frac{\alpha_i^{\overline{n}}\, \alpha_j^{\overline{m}}}{\alpha_0^{\overline{n+m}}} \left(\overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_i^{n,m}\right]}\overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_j^{m,n}\right]}\right.\\
&\qquad \left. \overline{\mathrm{Cov}[\log \hat{\mathbf{p}}_i^{n,m}, \log \hat{\mathbf{p}}_j^{n,m}]}\right) && (34)\\
&\quad + \frac{\alpha_i^{\overline{n}}\, \alpha_j^{\overline{m}}}{\alpha_0^{\overline{n}}\, \alpha_0^{\overline{m}}} \overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_i^{n}\right]}\overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_j^{m}\right]} \quad \textit{for } i \neq j\textit{, and}\\
&\mathrm{Cov}[\mathbf{p}_i^n \log \mathbf{p}_i, \mathbf{p}_i^m \log \mathbf{p}_i] && (35)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\overline{\alpha_i^{n+m}}}{\overline{\alpha_0^{n+m}}} \Big(\overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_i^{n+m}\right]}^2 \\
&\quad + \overline{\mathrm{Cov}[\log \hat{\mathbf{p}}_i^{n+m}, \log \hat{\mathbf{p}}_i^{n+m}]} \Big) \\
&\quad + \frac{\overline{\alpha_i^n}\,\overline{\alpha_i^m}}{\overline{\alpha_0^n}\,\overline{\alpha_0^m}} \overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_i^n\right]} \overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_j^m\right]}.
\end{aligned} \tag{36}
$$

Proof of Lemma D.3. This proof applies the well-know formula **(cov)** $\mathrm{Cov}[X, Y] = \mathbb{E}[X\,Y] - \mathbb{E}[X]\,\mathbb{E}[Y]$ once forward and once backward **(rcov)** $\mathbb{E}[X\,Y] = \mathrm{Cov}[X, Y] + \mathbb{E}[X]\,\mathbb{E}[Y]$ while applying Equation (10) several times:

$$
\begin{aligned}
&\mathrm{Cov}[\mathbf{p}_i^n \log \mathbf{p}_i, \mathbf{p}_j^m \log \mathbf{p}_j] && (37)\\
&\overset{\text{cov}}{=} \mathbb{E}\left[\mathbf{p}_i^n \log(\mathbf{p}_i)\, \mathbf{p}_j^m \log(\mathbf{p}_j)\right] \\
&\quad - \mathbb{E}\left[\mathbf{p}_i^n \log \mathbf{p}_i\right] \mathbb{E}\left[\mathbf{p}_j^m \log \mathbf{p}_j\right] && (38)\\
&= \frac{\overline{\alpha_i^n}\,\overline{\alpha_j^m}}{\overline{\alpha_0^{n+m}}} \mathbb{E}\left[\log(\hat{\mathbf{p}}_i^{i,j}) \log(\hat{\mathbf{p}}_j^{i,j})\right] \\
&\quad - \mathbb{E}\left[\log \hat{\mathbf{p}}_i^i\right] \mathbb{E}\left[\log \mathbf{p}_j^j\right] && (39)\\
&\overset{\text{(rcov)}}{=} \frac{\overline{\alpha_i^n}\,\overline{\alpha_j^m}}{\overline{\alpha_0^{n+m}}} \Big(\mathrm{Cov}[\log \hat{\mathbf{p}}_i^{i,j}, \log \hat{\mathbf{p}}_j^{i,j}] \\
&\qquad + \mathbb{E}\left[\log \hat{\mathbf{p}}_i^{i,j}\right] \mathbb{E}\left[\log \hat{\mathbf{p}}_j^{i,j}\right] \Big) && (40)\\
&\quad - \frac{\overline{\alpha_i^n}\,\overline{\alpha_j^m}}{\overline{\alpha_0^n}\,\overline{\alpha_0^m}} \mathbb{E}\left[\log \hat{\mathbf{p}}_i^i\right] \mathbb{E}\left[\log \mathbf{p}_j^j\right],
\end{aligned}
$$

where $\mathbf{p}^{i,j} \sim \mathrm{Dir}(\alpha^{i,j})$ with $\alpha^{i,j} = (\ldots, \alpha_i + n, \ldots, \alpha_j + m, \ldots)$. $\mathbf{p}^{i/j}$ and $\alpha^{i/j}$ are defined analogously. Applying Equation (9) and Equation (8) from Lemma D.1 yields the statement. For $i = j$, the proof follows the same pattern. □

D.3.2. Variance of Softmax Entropies

Now, we can prove the theorem that quantifies the variance of the entropy of Y:

Theorem D.5. *Given a Dirichlet distribution and a random variable* $\mathbf{p} \sim Dir(\alpha)$*, the variance of the entropy* $\mathrm{Var}_{\mathbf{p}\sim\mathrm{Dir}(\alpha)}\, \mathbb{H}_{Y\sim\mathrm{Cat}(\mathbf{p})}[Y]$ *of the categorical distribution* $Y \sim \mathrm{Cat}(\mathbf{p})$ *is given by*

$$
\begin{aligned}
&\mathrm{Var}[\mathbb{H}[Y \mid \mathbf{p}]] && (41)\\
&= \sum_i \frac{\overline{\alpha_i^2}}{\overline{\alpha_0^2}} \left(\overline{\mathrm{Cov}[\log \hat{\mathbf{p}}_i^2, \log \hat{\mathbf{p}}_i^2]} + \overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_i^2\right]}^2 \right) \\
&\quad + \sum_{i \neq j} \frac{\alpha_i\,\alpha_j}{\overline{\alpha_0^2}} \left(\overline{\mathrm{Cov}[\log \hat{\mathbf{p}}_i^1, \log \hat{\mathbf{p}}_j^1]} + \overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_i^{1,1}\right]}\, \overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_j^{1,1}\right]} \right) && (42)\\
&\quad - \sum_{i,j} \frac{\alpha_i\,\alpha_j}{\alpha_0^2} \overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_i^1\right]} \overline{\mathbb{E}\left[\log \hat{\mathbf{p}}_j^1\right]}.
\end{aligned}
$$

Proof. We start by applying the well-known formula $\mathrm{Var}[\sum_i X_i] = \sum_{i,j} \mathrm{Cov}[X_i, X_j]$ and then apply Lemma D.3 repeatedly. □