Calibrated Top-1 Uncertainty estimates for classification by score based models

Adam Oberman [1] **Chris Finlay** [1] **Alexander Iannantuono** [1] **Tiago Salvador** [2]

Abstract

While the *accuracy* of modern deep learning models has significantly improved in recent years, the ability of these models to generate *uncertainty estimates* has not progressed to the same degree. Uncertainty methods are designed to provide an estimate of class probabilities when predicting class assignment. While there are a number of proposed methods for estimating uncertainty, they all suffer from a lack of *calibration*: predicted probabilities can be off from empirical ones by a few percent or more. By restricting the scope of our predictions to only the probability of Top-1 error, we can decrease the calibration error of existing methods to less than one percent. As a result, the scores of the methods also improve significantly over benchmarks.

1. Introduction

Machine learning classification models have made groundbreaking gains in accuracy in recent years, using deep neural networks. However, our ability to predict when models make errors has lagged behind. As models are increasingly deployed to make decisions with real-world impacts, it becomes increasingly important to assess model error.

When machine learning models make a probabilistic prediction, the accuracy of the prediction must be assessed. For example if a model predicts that an image is dog with confidence .8, then 80% of this time this occurs, the image should be a dog. Any gap between predicted and empirical probabilities is referred to as *calibration* error. Methods which predict probabilities for class membership, can have significant calibration errors, even in the simplest setting of in-distribution error prediction.

*Equal contribution [1]Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada [2]Department of Mathematics, University of Michigan, Ann Arbor, Michigan, USA. Correspondence to: Adam Oberman <adam.oberman@mcgill.ca>.

Presented at the ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning. Copyright 2020 by the author(s).

1.1. Contributions

By focusing only on predicting the probability that the classification is *correct* on a given image, (Top-1), instead of predicting probabilities for all class labels, we are able to significantly reduce the calibration error of models, and improve the Brier scores compared to benchmark methods (Snoek et al., 2019). Calibration error is small, as illustrated in Figure 1. Table 5 shows the the improvements to scores using Top-1 binning: the expected calibration error (ECE) is reduced by a factor of three, and the Brier score is improved by a factor of three on ImageNet, and three to five on CIFAR-10. These improvements come from a better calibrated use of existing methods, rather than developing new methods.

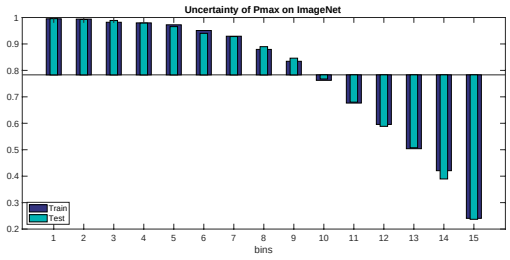


Figure 1. Predictions that image classification is correct (Top-1) using the pre-trained ResNet-152 model the ImageNet dataset. The values of Pmax are binned, based on the value of $Pmax$ uncertainty method (Hendrycks & Gimpel, 2017). Bins are designed to be equal size, and heights are plotted against the baseline Top-1 test accuracy of the model of 78.3%. The prediction is that images in the first bin will be correctly classified probability $.995 \pm .0025$, whereas images in the last bin will be correct with probability $.211 \pm .005$. Test and training bins are superimposed, with height gap indicating the difference between test and training predictions. The Expected Calibration Error (ECE), which measures the average gap, is 0.69%, which is significantly less than benchmark of 2-3% (Snoek et al., 2019).

2. Background

2.1. Classification and label probabilities

Let $y \in \mathcal{Y} = \{1, \dots, K\}$ be a label from one of K classes. Consider a model $f(x) = (f_1(x), \dots f_K(x))$

Table 1. Comparing the proposed Top-1 implementation of uncertainty measures with the benchmark implementation of the same measures. Top: Brier score (smaller is better) versus benchmarks in (Snoek et al., 2019). The Top-1 binning method scores better for each uncertainty measure. Bottom: Expected Calibration Error (ECE) for Top-1, compared to the range of ECE for methods in (Snoek et al., 2019).

BRIER SCORE	CIFAR-10		IMAGENET-1K	
UNCERTAINTY MEASURE	TOP-1	(SNOEK ET AL., 2019)	TOP-1	(SNOEK ET AL., 2019)
VANILLA	**0.033**	0.18	**0.11**	0.34
DROPOUT	**0.036**	0.17	**0.135**	0.35
ENSEMBLE VARIANCE	**0.040**	0.11	**0.135**	0.31

EXPECTED CALIBRATION ERROR	CIFAR-10	IMAGENET
TOP-1	**1.-1.1%**	**0.9-1.2%**
(SNOEK ET AL., 2019)	1.0-5.0%	2.0-3.0%

which outputs a score function f_i for each class i. The classification of the model corresponds to the highest score $\hat{y}(x) = \arg\max f_i(x)$. We say the model is Top-1 correct is $y = \hat{y}(x)$. Predictive uncertainty seeks to estimate the probability that an image belongs to each class,

$$p_k^{class}(x) = \mathbb{P}\left[y = k \mid x\right], \qquad \text{for each } k \in \mathcal{Y} \tag{1}$$

The softmax function transforms model outputs into a probability vector, $p_i^f(x) = \text{softmax}(f(x))_i$. The softmax values p^f are sometimes confused with $p_i^{class}(x)$, but generally speaking, $p_i^f(x)$ is not an accurate prediction of $p_i^{class}(x)$ (Domingos & Pazzani, 1996).

Consider instead to the Rank-k event, that the k-th largest score function corresponds to the correct label. Let $\hat{y}_k^{rank}$ be the index of k-th largest score function, and consider the probability of Rank-k,

$$p_k^{rank}(x) = \mathbb{P}\left[y = \hat{y}_k^{rank} \mid x\right], \qquad \text{for each } k \in \mathcal{Y} \tag{2}$$

In this work, we focus on estimating the Top-1 probability (which is the same as p_1^{rank}), which we write as $p^{TOP} = p_1^{rank}$. In the case of ImageNet, we also estimate and $p_5^{TOP} = \sum_{i=1}^{k} p_k^{rank}$. We estimate the probability using simple binning methods, described below. Our approach is a special case of the method in (Kuleshov & Liang, 2015), which sought to improve prediction errors by focusing on predicting only events of interest, in the context of structured prediction and image classification.

2.2. Predictive uncertainty methods

In the setting of deep neural networks, there are a number of proposed approaches for estimating (1). We refer to the recent work (Snoek et al., 2019) which compares a number of uncertainty methods, including: (i)(Vanilla) Maximum softmax probability (PMax) (Hendrycks & Gimpel, 2017) (ii) (Dropout) (Gal & Ghahramani, 2016) (iii) (Ensembles) Ensembles of M networks trained independently on the entire dataset using random initialization (Lakshminarayanan et al., 2017) (iv) (Temp Scaling) Post-hoc calibration by temperature scaling using a validation set (Guo et al., 2017).

Our implementation of these methods involves (i) choosing a scalar (single numeric) quantity $U(x)$ coming from the choice of method, and then binning the values. PMax is already a scalar value. For dropout, or an ensemble method, which return M vectors $f_1, \dots f_M$, the usual method is to return the variance of each component, as way to predict $p_i(xX)$. In our implementation, we combine the output into a single number: the variance of the norms of each of the vectors, $\text{Var}(\|f_i\|_2)$. Uncertainty results for temperature scaling are the same as for the original (unscaled) model, so we do not report these results. The reason is that temperature scaling a transformation of the output of the model which does not change the bins for our method.

Binning (Zadrozny & Elkan, 2001) is a simple method for predictive uncertainty. Instead of learning a parametric model to estimate (1), the values are simply measured empirically on a test set. Binning calibration error comes from to statistical error in bin probabilities in the test/training sets. These error can be significant if bin sizes are small, for example when there are many classes. Histogram binning as implemented in (Guo et al., 2017) led to a loss of accuracy in the multiclass setting. (Zadrozny & Elkan, 2002) separates multiclass problem into a number of binary problems, and calibrated the scores from each binary classifier and combine them to obtain multiclass probabilitiees. (Naeini et al., 2015) extends histogram binning to consider multiple binning models and their combination.

3. Calibrated Top-1 Uncertainty bins

We restrict the scope of our predictions to just the event that the classification of the model is *correct*, in other words, the Top-1 error, (2), or the Top-5 error.

Write $TOP(x) = 1_{\hat{y}=y}$ the Top-1 random variable ($TOP(x) = 1$ if the classification of the model is correct, and 0 otherwise). Let U be an uncertainty random variable. We define the conditional probability coming from a binning process as a random variable.

Definition 1 (Uncertainty bins)**.** Given the uncertainty random variable $U(x)$, and the edges $E = \{e_0, \dots, e_M\}$, define the prediction random variable, $C = (w, c)$ as the conditional probability

$$p_i^{TOP} = \mathbb{E}\left[TOP(x) \mid U(x) \in B_i\right] \tag{3}$$

with weights $w_i = \mathbb{P}\left[U \in B_i\right]$. Write $\bar{p} = \mathbb{E}\left[TOP\right]$ for the accuracy of the model.

The random variable C is a conditional probability, so it is calibrated, by definition (3). The empirical uncertainty predictions are empirical estimates of C Given the bin edges E, as above, and a data set,

- Evaluate U and TOP and bin the predictions of the values of $U(x)$ using the edges E.
- Let $\hat{w}_i$ be the fraction of points in bin i. The predicted confidence of the bin is given by

$$\hat{p}_i = \frac{1}{|B_i|} \sum_{j \in B_i} TOP(x_j)$$

We write $C^{\text{TRAIN}} = (w_i^{\text{TRAIN}}, p_i^{\text{TRAIN}}), C^{\text{TEST}} = (w_i^{\text{TEST}}, p_i^{\text{TEST}})$ for the test and training estimates of the prediction variable C.

More sophisticated estimates of the conditional probability could be performed using kernel density estimation, with bound for the probability density, as in (Wasserman, 2006).

3.1. Confidence intervals

Fix a bin, B_i, with $|B_i|$ elements, Then the empirical error $|p_i^{\text{TRAIN}} - p_i|$ is of order $1/\sqrt{|B_i|}$. More precisely, for any $\delta > 0$,

$$|p_i - p_i^{\text{TEST}}| \leq \sqrt{\frac{\log(2/\delta)}{2|B_i|}} \qquad \text{with probability } \geq 1 - \delta \tag{4}$$

The result follows from Hoeffding's inequality in the case of Bernoulli random variables (Wasserman, 2006).

4. Scoring the predictions

4.1. Proper scoring rules: Brier and NLL

Proper scoring rules (Gneiting & Raftery, 2007) (Parmigiani & Inoue, 2009) are used to score predictive uncertainty estimates. In our setting, calibration error measures how far empirical bin probabilities $\hat{p}_i$ are from the true values p_i, given by (3). Resolution measures how far the values p_i are from the accuracy $\bar{p}$. Scoring rules can be decomposed into terms involving calibration error and resolution, see §C.

In the binary case, the Brier score (Brier, 1950) is given by

$$BS(TOP, C) = \mathbb{E}\left[(p(x) - TOP(x))^2\right]$$

In this case, the resolution is the variance of C, and the calibration error is the $\sum w_i |p_i - \hat{p}_i|^2$.

The negative Log-Likelihood (NLL) score (Good), $NLL(TOP, C)$, is given by

$$-\mathbb{E}\left[\log(p(x))TOP(x) + \log(1 - p(x))(1 - TOP(x))\right]$$

4.2. Expected Calibration error

The Expected Calibration error (ECE) is a popular method for measuring calibration error. The ECE between an empirical prediction variable $\widehat{C}$ and the true prediction random variable C defined by (3) is given by

$$ECE(C, \widehat{C}) = \sum_i w_i |p_i - \hat{p}_i|.$$

The ECE is approximated using empirical test and training histograms by

$$ECE(C^{\text{TEST}}, C^{\text{TRAIN}}) = \sum_i w_i^{\text{TEST}} |p_i^{\text{TEST}} - p_i^{\text{TRAIN}}|.$$

A large ECE is bad, but a zero ECE is not necessarily good. For example, always predicting $\bar{p}$ gives an ECE of zero but has zero resolution.

4.3. Odds ratio as a measure of resolution

Given a probability $p \in (0, 1)$, the odds of p, is given by $O(p) = p/(1 - p)$. We propose the odds ratio as a measure of resolution.

Definition 2 (Odds Ratio)**.** For $p, a \in (0, 1)$ define

$$OR(p, a) = \max\left(\frac{O(p)}{O(a)}, \frac{O(a)}{O(p)}\right) \tag{5}$$

The Expected Odds Ratio is $EOR(C) = \sum_i w_i OR(p_i, \bar{p})$. The EOR measures the resolution of C in terms of the amount we expect our winnings to increase, is we could make bets with odds $O(\bar{p})$, given the information provided the probability bins. We present a motivating example below in §C.1.

5. Empirical results

We discuss details of the implementation in §A.1.

5.1. Uncertainty on CIFAR-10

We compare different uncertainty methods on CIFAR-10: Pmax, Dropout, and data from CIFAR-10H (Peterson et al., 2019). On the human classified images, we used the entropy of the human probability vector for each image. We found that dropout with variance .02 was competitive, but dropout with variance .005 was not effective. Human scoring was better than the weak dropout, but not as effective as the other methods.

The scores are presented in Table 2 and visualized in Figure 2. We compared results using 10 bins and 100 bins. These results confirm the notion that our models are calibrated, up to error coming from bin sampling, which is small. The calibration error is small for all models, with ECE of 0.0096 ± 0.0029 for 10 bins, and ECE at most $.020 \pm .002$ for 100 bins.

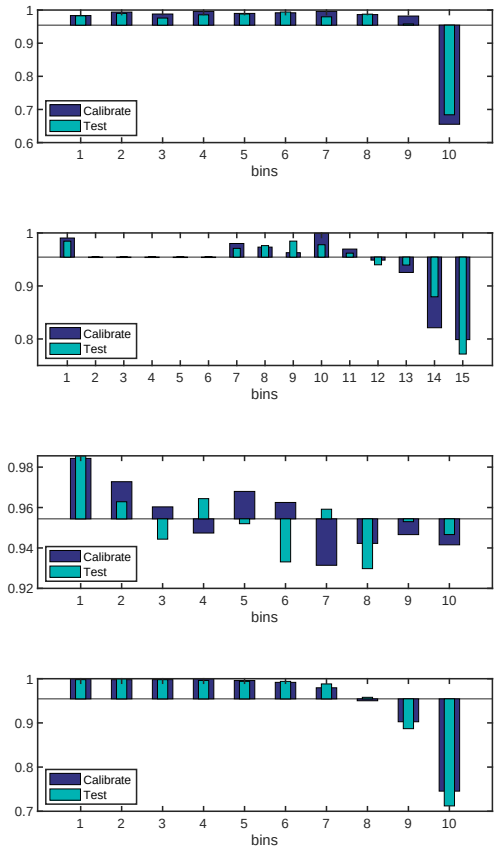


Figure 2. Confidence bins on CIFAR-10: PMax, Human classification, Dropout .005, Dropout .02. The first dropout is not effective: the bins are small and poorly calibrated.

Table 2. Scoring on CIFAR-10 using 10 bins (and 100 bins)

	BRIER	NLL	EOR
VANILLA	**0.0343**	**0.13**	$5.51 \pm .8$
HUMAN	0.0394	0.16	$3.02 \pm .4$
DROPOUT .005	0.0436	0.18	$1.42 \pm .01$
DROPOUT .02	0.0378	**0.13**	$\mathbf{98.7 \pm 66}$
VANILLA (100 BINS)	**0.0343**	**0.14**	26 ± 3
HUMAN (100 BINS)	0.0406	0.16	4 ± 1.4
DROPOUT .005 (100 BINS)	0.0445	0.20	7 ± 1
DROPOUT .02 (100 BINS)	0.0372	**0.14**	$\mathbf{32 \pm 1.5}$

Table 3. Scoring Top-1 on ImageNet using 20 bins (100 bins)

	BRIER	NLL	EOR
VANILLA	**0.115**	**0.358**	$\mathbf{13 \pm 2}$
ENTROPY	0.117	0.365	12 ± 1
DROPOUT .005	0.168	0.518	$1.3 \pm .02$
DROPOUT .02	0.136	0.408	8.9 ± 0.8
VANILLA (100 BINS)	**0.116**	**0.361**	$\mathbf{21 \pm 2}$
ENTROPY (100 BINS)	0.118	0.368	20 ± 3
DROPOUT .005 (100 BINS)	0.168	0.516	$1.5 \pm .02$
DROPOUT .02 (100 BINS)	0.136	0.411	14 ± 4

Table 4. Scoring Top-5 on ImageNet using 20 bins (100 bins)

	BRIER	NLL	EOR
VANILLA	0.048	0.175	$\mathbf{20 \pm 20}$
ENTROPY	**0.046**	**0.166**	9 ± 1
DROPOUT .005	0.056	0.223	$1.3 \pm .02$
DROPOUT .02	0.053	0.194	7 ± 1
VANILLA (100 BINS)	0.047	0.176	$\mathbf{35 \pm 5}$
ENTROPY (100 BINS)	**0.046**	**0.168**	$\mathbf{32 \pm 5}$
DROPOUT .005	0.056	0.227	2 ± 1
DROPOUT .02 (100 BINS)	0.052	0.195	27 ± 6

5.2. Top-1 and Top-5 uncertainty on ImageNet

In this section, we score the Top-5 and Top-1 accuracy on ImageNet, comparing results with 20 and 100 bins. See Tables 3 and 4 and Figure 3

Calibration: the ECE was $.01 \pm .002$ ($.02 \pm .002$) for three of the measures, and $0.012 \pm .002$ ($0.03 \pm .002$) for the less effective dropout, using 20 and 100 bins, respectively. The Brier and ECE scores were consistent across both 20 and 100 bins, which suggests that the increase in resolution was balanced by the decreased calibration. The EOR increased when the number of bins was increases, as expected, since it is a measure of spread that does not take into account calibration.

References

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, January 1950. doi: 10.1175/1520-0493(1950)078⟨0001:vofeit⟩2.

0.co;2.

Bröcker, J. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 135 (643):1512–1519, 2009.

Davis, J. and Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

DeGroot, M. H. and Fienberg, S. E. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Domingos, P. and Pazzani, M. Beyond independence: Conditions for the optimality of the simple bayesian classi er. In *Proc. 13th Intl. Conf. Machine Learning*, pp. 105–112, 1996.

Gal, Y. and Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pp. 1050–1059, 2016. URL http://proceedings.mlr.press/v48/gal16.html.

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

Good, I. J.(1952). rational decisions. *Journal of the Royal Statistical Society, Ser. B*, pp. 107–114.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 1321–1330, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

Hendrycks, D. and Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Kuleshov, V. and Liang, P. Calibrated structured prediction. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 3474–3482, 2015. URL http://papers.nips.cc/paper/5658-calibrated-structured-prediction.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 6405–6416, 2017.

Naeini, M. P., Cooper, G. F., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pp. 2901–2907, 2015. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9667.

Parmigiani, G. and Inoue, L. *Decision theory: Principles and approaches*, volume 812. John Wiley & Sons, 2009.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., and Russakovsky, O. Human uncertainty makes classification more robust. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9617–9626, 2019.

Pleiss, G., Zhang, T., Elenberg, E. R., and Weinberger, K. Q. Identifying mislabeled data using the area under the margin ranking. *arXiv preprint arXiv:2001.10528*, 2020.

Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., and Rabinovich, A. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

Snoek, J., Ovadia, Y., Fertig, E., Lakshminarayanan, B., Nowozin, S., Sculley, D., Dillon, J. V., Ren, J., and Nado, Z. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 13969–13980, 2019.

Wasserman, L. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

Zadrozny, B. and Elkan, C. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 204–213, 2001.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, pp. 694–699, 2002. doi: 10.1145/775047.775151. URL https://doi.org/10.1145/775047.775151.

A. Supplementary Material

A.1. Methodology

In practice, we start with a calibration (test) set, choose the number of bins, M, and define the bin edges so that there are equal number of points in each bin. We chose bin sizes so that our ECE was always less that one percent, which in practice was consistent with the $1/\sqrt{|B_i|}$ estimate above. We chose to calibrate using half the test set, and measure the error using the other half. 10-fold cross validation lead to ECE of $< 1.\%$ with a standard deviation $< 0.1\%$ using on the order of 10-20 bins. At 100 bins, the ECE was slightly over 1%. The choice of bin size did not affect the scores too much: the change to the scores was less that the differences of scores between different methods. Thus ECE, scoring, and relative ranking of method are insensitive to bin sizes.

In order to avoid probabilities of 0, or 1, we regularized the probabilities as follows $\hat{p}_i^{\epsilon} = \frac{|B_i|\hat{p}_i + p_A}{|B_i|+1}$, where p_A is the accuracy of the model. This regularization, which corresponds to adding a single element to the bin with value p_A, can be interpreted as using a Bayesian prior with mean p_A for the bin. The regularization made a difference in the case of the Cifar-10H (human) data which otherwise had empty bins. The regularization did not make a difference for the other methods, where bin probabilities were already between 0 and 1.

We have enough data so that the bin sizes allow us to use the bound (4) to obtain useful estimates on the error in our predictions. For example, in the results which follow in §5.1 combining the first five (of ten total) bins for Dropout .02, we obtain $p_1^{\text{TEST}} = 0.9983$ with $|B_1| = 2500$ which leads to a useful confidence interval. For example, taking $\delta = .005$ we get a bound of .034 for the error in the estimate with probability at least .995.

B. Additional Empirical Results

B.1. Discussion of CIFAR-10 results

In most cases, is little difference in the scores going from 10 to 100 bins, on the order of 10% relative change. However, for EOR, the scores increase going from 10 to 100 bins, which is expected, since we are resolving more fine grained probabilities. So, for example, the Vanilla method went from 5.5 to 26 on EOR. Dropout .02, which has the highest EOR value was not resolved at 10 bins: the relative deviation in the EOR was order one. However, at 100 bins, the score stabilized to the high value of 32 with standard deviation of 1.5.

Across both 10 and 100 bins, the Brier score ranks (Vanilla) best, the NLL score ranks (Vanilla) and Dropout .2 equally. According to EOR, the Dropout .2 is best, but the variance is large for 10 bins. At 100 bins, the variance of EOR is small. The difference in the scores is explained by looking at the bin probabilities. For Pmax the first five bins have average value of 0.9886, with corresponding odds ratio of 87. For Dropout, the first five bins have an average value: 0.9983, with corresponding odds ratio of 592. It is the higher probabilities in the Dropout bins that contribution to the higher EOR score. Thus, in the regime of very accurate models, we should use EOR to make fine-grained distinctions between calibrated predictive uncertainty measures, although we should also watch the variance of the EOR.

B.2. Discussion Top-1 and Top-5 uncertainty on ImageNet

Refer to Tables 3 and 4 and Figure 3

For Top-1 accuracy, Vanilla was the best method, according to Brier, NLL, and EOR. Dropout .005 was the lease effective measure, as illustrated by the EOR, which is close to the minimum of 1. For Top-5 accuracy, Entropy comes out at top ranked by both Brier and NLL. However the EOR spread is better for Vanilla with 10 bins, and nearly equal for Vanilla and Entropy with 100 bins. It is reasonable the Vanilla worked best for Top-1, since it corresponds to PMax, likewise, entropy also takes into account the lower ranked probabilities, so it seems reasonable that it works better for Top-5.

In terms of scoring methods, NLL had a wider relative range of values, compared to the Brier score, so it made larger distinction between methods, but the ranking of the methods is the same. EOR can be inconsistent for highly accurate models with small number of bins, however the wide range of scores emphasizes distinctions between the models.

C. Scoring Discussion

C.0.1. Brier and NLL scores decomposed

(DeGroot & Fienberg, 1983) studies probabilistic scoring methods in the binary case, in particular the decomposition of scoring rules into calibration error and resolution. This decomposition was generalized to multiclass (Bröcker, 2009).

Let C be the exact prediction random variable and let $\widehat{C}$ be an empirical prediction random variable. For the Brier

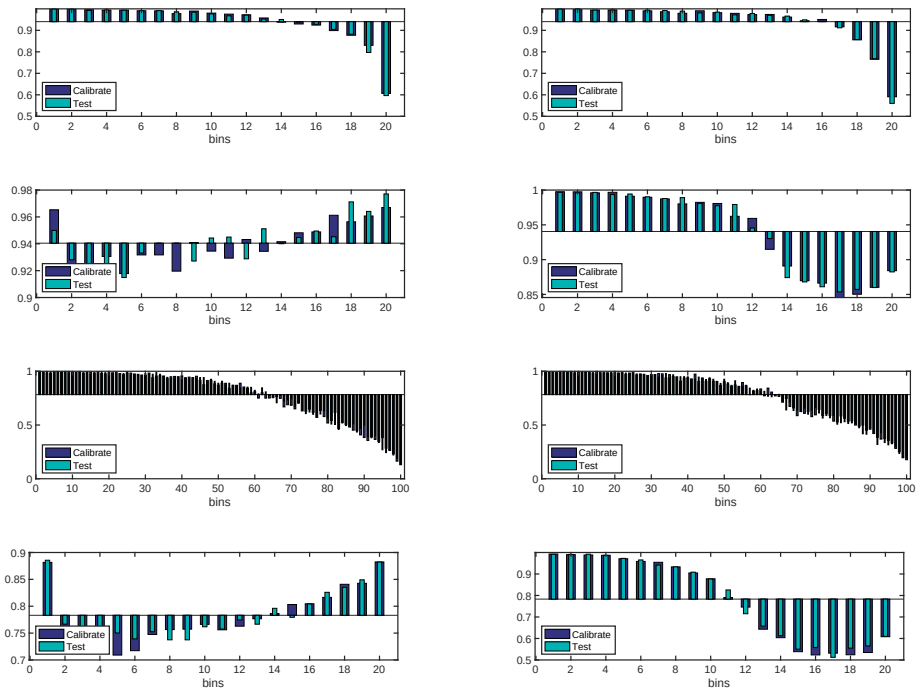


Figure 3. Top-5 (first four) and Top-1 (next four) Confidence bins on ImageNet In order from top left to bottom right: Vanilla, Entropy, Dropout .02, Dropout .005. Note the shape of the bins is similar, but the width depends on the accuracy.

score, we have the decomposition

$$BS(TOP, \widehat{C}) = \underbrace{\bar{p}(1-\bar{p})}_{\text{uncertainty}} - \underbrace{\sum w_i(\bar{p}-p_i)^2}_{\text{resolution}} + \underbrace{\sum w_i(p_i-\hat{p}_i)^2}_{\text{calibration error}}$$

The resolution is just the variance of the accuracy over the bins, which is the width of the centered bins in Figure 1. The calibration error is the mean-squared difference of confidence and accuracy over the bins. Thus the Brier score is minimized if (i) calibration error is zero and (ii) the resolution is as large as possible (equal to the uncertainty).

The decomposition of the NLL score is given by

$$NLL(TOP, \widehat{C}(x)) = \underbrace{H(\bar{p})}_{\text{uncertainty}} - \underbrace{\sum w_i KL(\bar{p}, p_i)}_{\text{resolution}} + \underbrace{\sum w_i KL(\hat{p}_i, p_i)}_{\text{calibration error}}$$

where the entropy is $H(p) \sum -p_i \log(p_i)$ and the divergence term is given by the KL divergence $KL(p,q) = -\sum q_i \log(p_i/q_i)$. Again, we have an interpretation of the terms: entropy is an information theoretical measure of the uncertainty, and the KL divergence terms are measures of spread.

The derivation of the decomposition of the Brier score is as follows (see, for example, (Kuleshov & Liang, 2015))

$$\begin{aligned} & BS(TOP, \widehat{C}) \\ &= \mathbb{E}\left[(TOP - \widehat{C})^2\right] \\ &= \mathbb{E}\left[(TOP - C)^2\right] + \mathbb{E}\left[(C - \widehat{C})^2\right] \\ &= \underbrace{\text{Var}\left[TOP\right]}_{\text{uncertainty}} - \underbrace{\text{Var}\left[C(x)\right]}_{\text{resolution}} + \underbrace{\mathbb{E}\left[(\widehat{C}(x) - C(x))^2\right]}_{\text{calibration error}} \end{aligned}$$

C.1. Explaining the Odds Ratio resolution

Example 1. Consider the experiment consisting of tossing one of four randomly chosen coins. Coins 1 and 2 are fair, while coins 3 and 4 have $p(H) = 15/16$ and $1/16$, respectively. Clearly, $p(H) = .5$ for this experiment. Next, suppose we can identify the coins, and let $U \in \{1, 2, 3, 4\}$ represent the chosen coin. Conditioning on the value of U, we have the following histogram: $p = (1/2, 1/2, 15/16, 1/16)$, $w = (.25, .25, .25, .25)$. Knowing which coin is being tossed changes the odds of heads from 1:1 to 15:1 or to 1:15.

The odds for the baseline probability are $O(a) = a/(1 -$

a). If new information changes the probability to p, the new odds are $O(p) = p/(1-p)$. The value of the new information must be compared to the odds. The odds ratio measures how much the expected winnings of a fair bet increase, when new information is available. When $p \geq a$ are given by the odds ratio is $OR(p,a) = \frac{O(p)}{O(a)}$. On the other hand, if the odds have decreased, we should bet against, so the odds ratio is $OR(p,a) = \frac{O(a)}{O(p)}$.

The EOR for the coin toss Example 1 is 8. This value corresponds to the fact that for a single bet, we expect to win 15 each time the biased coins are tossed, and 1 each time the fair coins, so the EOR is (15 + 15 + 1 + 1)/4 = 8.

D. More scoring

D.1. Comparison of the expected odds ratio and Brier score

We compared the expected odds ratio (EOR) and Brier score for various uncertainty methods on benchmark image classification datasets. Since our methods are well-calibrated, we can see if the indication of resolution coming from the EOR gives more information.

In Table 5, we present the Brier scores (smaller is better) for uncertainty methods on In Table 6 we show the EOR (larger is better) for various uncertainty methods, on CIFAR-10, CIFAR-100, and ImageNet-1K.

On ImageNet, the model based uncertainty metrics perform better than dropout variance and ensemble variance, which are more expensive to compute, with regards to both Brier score and EOR. On the other hand, on CIFAR-10, Brier score prefers model entropy and Pmax, whereas the EOR prefers Ensemble variance and the best choice of dropout. Finally, on CIFAR-100, the rankings are different again : ensemble variance has the best Brier score, and comes in second for EOR. Entropy, Pmax and the $-\log(\sum p_{1:5})$ are exactly equal for the Brier score, but differently ranked by EOR, with the third one being best for EOR.

D.1.1. Finding mislabelled images

We used Pmax to detect incorrectly labelled images see (Reed et al., 2014) and (Pleiss et al., 2020) . We looked for the most confident images which were incorrectly classified, using the most confident bins. It turned out that all of these were either incorrectly labelled, or were ambiguous images, see illustrations in Figure 4. For example, in the second image, the animal is a wallaby, not a wombat. In the fourth image, a paintbrush is a kind of plant, but there is also a pot in the image.

Table 5. Brier score of various measures of uncertainty. For CIFAR-10 we used X_1, the probability of the correct label; for CIFAR-100 and ImageNet-1K we used X_5 the probability that the correct label is in the Top5. Data is binned into 100 bins, chosen to have equal weight.

UNCERTAINTY MEASURE	CIFAR-10	CIFAR-100	IMAGENET-1K
MODEL ENTROPY	**0.033**	0.067	0.041
$-\log p_{\text{MAX}}$	**0.033**	0.067	0.042
$-\log \sum p_{1:5}$	-	0.067	**0.040**
DROPOUT ($p = 0.002$)	0.036	0.074	0.047
DROPOUT ($p = 0.01$)	0.04	0.075	0.048
DROPOUT ($p = 0.05$)	0.043	0.076	0.049
ENSEMBLE VARIANCE	0.040	**0.050**	0.047
LOSS	0	0.029	0.019

Table 6. Expected odds ratio for various uncertainty methods. For CIFAR-10 we used the probability of the correct label; for CIFAR-100 and ImageNet-1K we used the probability that the correct label is in the top 5. Data is binned into 100 bins, chosen to have equal weight.

UNCERTAINTY METHOD	CIFAR-10	CIFAR-100	IMAGENET
MODEL ENTROPY	4.29	3.64	8.18
$-\log p_{\text{MAX}}$	4.22	3.77	**8.87**
$-\log \sum p_{1:5}$	-	**4.25**	8.45
DROPOUT ($p = .002$)	10.39	3.11	6.84
DROPOUT ($p = .01$)	4.67	2.38	7.81
DROPOUT ($p = .05$)	1.69	1.35	1.60
ENSEMBLE VARIANCE	**16.66**	4.03	6.13

D.2. Metrics for discrimination: ROC and AUROC

The receiver operating characteristic curve (ROC) a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied (Davis & Goadrich, 2006). The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. This plot is summarized in a single statistic, which is the area under the ROC curve (AUROC), which is in the interval $(0, 1)$. The AUROC was used as a score for uncertainty in (Hendrycks & Gimpel, 2017)

The AUROC score requires the uncertainty random variable to have ordered bins, and forces the decision threshold to combine the probabilities from each bin. In example 2 we show that different orderings of the histogram lead to different AUROC values. In example 3 we show that the AUROC can be the same, for two examples with very different expected odds ratios, in particular, the second histogram has a very high accuracy bin which is reflected in the EOR, but not as much in the AUROC.

Example 2 (AUROC depends on ordering)**.** Consider the coin toss example 1. In this case, with the increasing or-



Figure 4. High uncertainty (low loss) images which are incorrectly classified. These are mislabelled or ambiguous.

dering of the probabilities the AUROC is .83. On the other hand with the reverse ordering, the AUROC is .17, and it can take other values in between. The expected odds ratio (EOR) score is 8, independent of the ordering.

Example 3 (same AUROC different BR)**.** Consider two equally weighted histograms given by $p = (.15, .4, .8)$, $q = (.4, .5, .99)$. The averages are different: $\bar{p} = .45, \bar{q} = .63$. The AUROC$(p, w)$ = .79, AUROC(q, w) = .78. EOR(p, w) = 3.6, EOR(q, w) = 21.

D.2.1. Scoring rules and refinement

In this question we show that refining a histogram increases the measures of spread, such as the variance, the

hat convex scoring rules can only increase if we refine the histogram. In particular, this applies to the expected odds ratio, by taking $S = OR$. This result means that having more data means we can refine the histogram, and, as a result, get a better expected score. This section is the most mathematical part of the paper, it can be skipped on first reading.

The histogram random variable U given by u_i, w, p, defined by (3). In order to prove the theorem, we need to mathematically define what we mean by a refined histogram, which is intuitively simple, but requires some notations. To begin, take a simple example where we split one bin into equal halves. When the bin i has values w_i, p_i is into two equal parts (so with weights $b = (.5, .5)$), then we obtain a new histogram, with w_i, p_i replaced by $(w_i/2, w_i/2), (p_1, p_2)$. Note that the p_1, p_2 will have different values, but the expectation must be the same: $(p_1 + p_2)/2 = p_i$. We generalize the idea of refining a histogram as follows.

Definition 3 (Refined histogram)**.** Begin with the histogram U and values p_i, w_i, for $i \in [n]$. The histogram U' with values p'_j, w'_j, $j \in [m]$ is a *refinement* of U if the indices $[m]$ are partitioned into subsets $J_1, \dots J_n$ so that each index i is identified with an index set J_i, and with a corresponding weight vector $b = b(i)$, with $\sum b_j = 1, b_j > 0$ such that

$$w'_j = b_j w_i, \qquad p_i = \sum_{j \in J} b_j p'_j \tag{6}$$

Note (6) is a generalization of the example of splitting one index into two parts.

Now we are prepared to prove the theorem, which will be a result of Jensen't inequality applied to a refined histogram random variable. Recall that for a convex function $f(x)$, and for a weight vector b, Jensen's inequality says

$$f\left(\sum_k b_k x_k\right) \le \sum_k b_k f(x_k) \tag{7}$$

Theorem 1. *Let U be a histogram random variable, and let U' be a refinement of U. Let $S(p, a)$ be any convex scoring rule. Then refining the histogram can only increase the expected score:*

$$\mathbb{E}S(X \mid U) \le \mathbb{E}S(X \mid U') \tag{8}$$

Proof. First, from the definition of expected scoring rule,

$$\mathbb{E}S(X \mid U) = \sum_i w_i S(p_i, a)$$

Fix one index, i, and let J_i and $b = b(i)$ be the corresponding index set and weight vector. Then

$$S(p_i, a) = S\left(\sum_{j \in J} b_j p'_j, a\right) \le \sum_{j \in J} b_j S(p'_j, a)$$

using the second part of (6), and using Jensen's inequality (7), since S is convex. Next, multiply by w_i to obtain

$$w_i S(p_i, a) \le \sum_{j \in J} w_i b_j S(p'_j, a)$$

using the first part of (6) gives

$$w_i S(p_i, a) \le \sum_{j \in J} w'_j S(p'_j, a)$$

next, summing over i gives $\mathbb{E}S(X \mid U)$ on the left hand side, and $\mathbb{E}S(X \mid U')$ on the right hand side, which is the desired result. □