Supplemental Material for "Active Bayesian Assessment of Black-Box Classifiers"

Disi Ji,[1] **Robert L. Logan IV,** [1] **Padhraic Smyth** [1] **Mark Steyvers** [2]
[1] Department of Computer Science, University of California, Irvine
[2] Department of Cognitive Sciences, University of California, Irvine
disij@uci.edu, rlogan@uci.edu, smyth@ics.uci.edu, mark.steyvers@uci.edu

Supplemental Material for Section 4: Bayesian Assessment

Bayesian Metrics

Bayesian Estimates of Reliability Diagrams One particular application of Bayesian groupwise accuracy estimation is to **reliability diagrams**. Reliability diagrams are a widely used tool for visually diagnosing model calibration (DeGroot and Fienberg 1983; Niculescu-Mizil and Caruana 2005). These diagrams plot the empirical sample accuracy $A_M(\mathbf{x})$ of a model M as a function of the model's confidence scores $s_M(\mathbf{x})$. If the model is perfectly calibrated, then $A_M(\mathbf{x}) = s_M(\mathbf{x})$ and the diagram consists of the identity function on the diagonal. Deviations from the diagonal reflect miscalibration of the model. In particular if the curve lies below the diagonal with $A_M(\mathbf{x}) < s_M(\mathbf{x})$ then the model M is overconfident (e.g., see Guo et al. (2017)). For a particular value $s_M(\mathbf{x}) = s \in [0, 1]$ along the x-axis, the corresponding y value is defined as: $\mathbb{E}_{\mathbf{x}|s_M(\mathbf{x})=s}[A_M(\mathbf{x})]$.

To address data sparsity, scores are often aggregated into bins. We use equal-width bins here, denoting the b-th bin or region as $\mathcal{R}_b = \{\mathbf{x}|s_M(\mathbf{x}) \in [(b-1)/B, b/B)\}$, where $b = 1, \ldots, B$ ($B = 10$ is often used in practice). The unknown accuracy of the model per bin is θ_b, which can be viewed as a marginal accuracy over the region $\mathcal{R}_b$ in the input space corresponding to $s(\mathbf{x}) \in \mathcal{R}_b$, i.e.,

$$\theta_b = \int_{\mathcal{R}_b} p(y = \hat{y}_M|\mathbf{x})p(\mathbf{x}|\mathbf{x} \in \mathcal{R}_b)d\mathbf{x}.$$

As described in the main paper, we can put Beta priors on each θ_b and define a binomial likelihood on outcomes within each bin (i.e., whether a model's predictions are correct or not on each example in a bin), resulting in posterior Beta densities for each θ_b.

In Figure 1 we show the Bayesian reliability diagrams for the five datasets discussed in the main paper. The columns indicate different datasets and the rows indicate how much data (from the test set) was used to estimate the reliability diagram. Based on the full set of test examples (row 3), the posterior means and the posterior 95% credible intervals are generally below the diagonal, i.e., we can infer with high confidence that the models are miscalibrated (and overconfident, to varying degrees) across all five datasets. For some bins where the scores are less than 0.5, the credible intervals are wide due to little data, and there is not enough information to determine with high confidence if the corresponding models are calibrated or not in these regions. With $N = 100$ examples (row 1), the posterior uncertainty captured by the 95% credible intervals indicates that there is not yet enough information to determine whether the models are miscalibrated given only $N = 100$ labeled examples. With $N = 1000$ examples (row 2) there is enough information to reliably infer that the CIFAR-100 model is overconfident in all bins for scores above 0.3. For the remaining datasets the credible intervals are generally wide enough to include 0.5 for most bins, meaning that we do not have enough data to make reliable inferences about calibration, i.e., the possibility that the models are well-calibrated cannot be ruled out without acquiring more data.

Bayesian Estimation of Calibration Performance (ECE) As shown in Figure 1, "self-confident" estimates provided by machine learning predictors can often be quite unreliable and miscalibrated (Zadrozny and Elkan 2002; Kull, Silva Filho, and Flach 2017; Ovadia et al. 2019). In particular, complex models such as deep networks with high-dimensional inputs (e.g., images and text) can be significantly overconfident in practice (Gal and Ghahramani 2016; Guo et al. 2017; Lakshminarayanan, Pritzel, and Blundell 2017).

We can assess calibration-related metrics for a classifier in a Bayesian fashion using any of well-known various calibration metrics which are defined as discrepancy between model score and accuracy (Kumar, Liang, and Ma 2019; Nixon et al. 2019). Here we focus on expected calibration error (ECE) given that it is among the widely-used calibration metrics in the machine learning literature (e.g., Guo et al. (2017); Ovadia et al. (2019)).

As discussed in the previous subsection, we use the standard ECE binning procedure, then the marginal ECE is defined as a weighted average of the absolute distance between the binwise accuracy θ_b and the average score s_b per bin:

$$\text{ECE} = \sum_{b=1}^{B} p_b|\theta_b - s_b| \tag{1}$$

where p_b is the probability of a score lying in bin b and it can

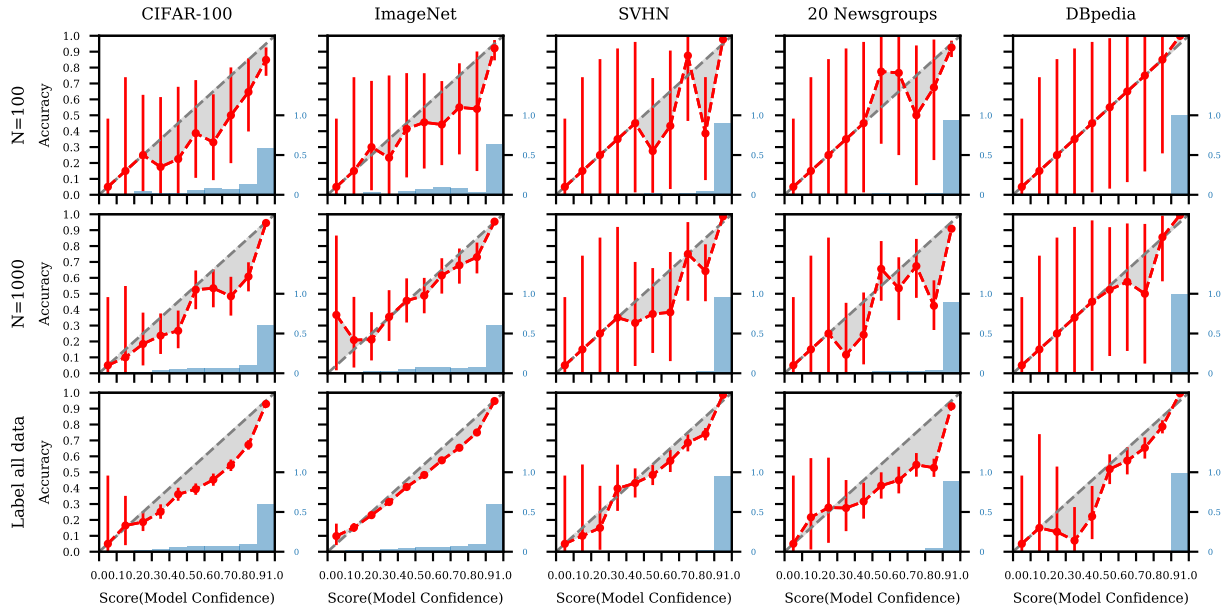


Figure 1: Bayesian reliability diagrams for five datasets (columns) estimated using varying amounts of test data (rows). The red circles plot the posterior mean for θ_j under our Bayesian model. Red bars display 95% credible intervals. Shaded gray areas indicate the estimated magnitudes of the calibration errors, relative to the Bayesian estimates. The blue histogram shows the distribution of the scores for N randomly drawn samples.

be estimated with all unlabeled data available. As shown in Figure 2, the weights of these bins tend to be quite skewed in practice for deep neural networks models like ResNet-110.

While the posterior of ECE is not available in closed form, its Monte Carlo samples are straightforward to obtain, by drawing random samples from Beta distributions of θ_b and deterministically computing ECE with Equation 1.

Figure 2 shows the results of Bayesian estimation of a reliability diagram (top row) and the resulting posterior estimate of ECE (bottom row) for the CIFAR-100 dataset with three different values of N. The third column for $N = 10000$ corresponds to using all of the data in the test set. The other 2 columns of plots correspond to particular random samples of size $N = 100$ and $N = 1000$. The ECE value computed using all the test data ($N = 10000$) is referred to as ground truth in all plots, "Bayesian" refers to the methodology described in the paragraphs above, and "frequentist" refers to the standard frequentist estimate of ECE.

The bottom row of Figure 2 plots empirical samples (in red) from the posterior density of ECE as the amount of data increases. As N increases the posterior of ECE converges to ground truth, and the uncertainty about ECE decreases. When the number of samples is small ($N = 100$), the Bayesian posterior for ECE puts non-negative probability mass on ground truth marginal ECE, while the frequentist method significantly overestimates ECE without any notion of uncertainty.

Figure 3 shows the percentage error in estimating ground truth ECE, for Bayesian mean posterior estimates (MPE) and frequentist estimates of ECE, as a function of the number of labeled data points ("queries") across the five datasets in the paper. The percentage is computed relative to the ground truth marginal ECE $=$ ECE*, computed with the whole test set as before. The MPE is computed with Monte Carlo samples from the posterior distribution (histograms of such samples are shown in Figure 2). At each step, with we randomly draw and label N queries from the pool of unlabeled data, and compute both a Bayesian and frequentist estimate of marginal calibration error with these labeled data. We run the simulation 100 times, and report the average ECE$_N$ over the N samples. Figure 3 plots $(\text{ECE}_N - \text{ECE}^*)/\text{ECE}^*$ as a percentage. The Bayesian method consistently has lower ECE estimation error, especially when the number of queries is small.

Bayesian Estimation of ECE per Class. Similar to accuracy, we can also model *classwise ECE*,

$$\text{ECE}_k = \sum_{b=1}^{B} p_{b,k} |\theta_{b,k} - s_b|$$

by modifying the model described above to use regions $\mathcal{R}_{b,k} = \{\mathbf{x} | \hat{y} = k, s(\mathbf{x}) \in \mathcal{R}_b\}$ that partition the input space by predicted class in addition to partitioning by the model score. This follows the same procedure as for "total ECE" in the previous subsection except that the data is now partitioned by predicted class $k = 1, \ldots, K$ and a posterior density on ECE$_k$ for each class is obtained.

In the main paper, we showed a scatter plot of classwise accuracy and classwise ECE assessed with our proposed Bayesian method for CIFAR-100. In Figure 4 we show scatter plots for all five datasets used in the paper. The assessment shows that model accuracy and calibration vary substantially across classes. For CIFAR-100, ImageNet and 20 Newsgroups, the variance of classwise accuracy and ECE

among all predicted class is considerably greater than the variance of two other datasets. Figure 4 also illustrates that there is significant negative correlation between classwise accuracy and ECE across all 5 datasets, i.e. classes with low classwise accuracy also tend to be less calibrated.

Bayesian Estimation of Confusion Matrices Conditioned on a predicted class $\hat{y}$, the true class label y has a categorical distribution $\theta_{jk} = p(y = j|\hat{y} = k)$. We will refer to θ_{jk} as confusion probabilities. In a manner similar to using a beta-binomial distribution to model accuracy, we can model these confusion probabilities using a Dirichlet-multinomial distribution:

$$\theta_{\cdot k} \sim \text{Dirichlet}(\alpha_{\cdot k}) \tag{2}$$

There are $\mathcal{O}(K^2)$ parameters in total in K Dirichlet distributions, each of which is parameterized with a K dimensional vector α_j.

Bayesian Misclassification Costs Accuracy assessment can be viewed as implicitly assigning a binary cost to model mistakes, i.e. a cost of 1 to incorrect predictions and a cost of 0 to correct predictions. In this sense, identifying the predicted class with lowest accuracy is equivalent to identifying the class with greatest expected cost. However, in real world applications, costs of different types of mistakes can vary drastically. For example, in autonomous driving applications, misclassifying a pedestrian as a crosswalk can have much more severe consequences than other misclassifications.

To deal with such situations, we extend our approach to incorporate an arbitrary cost matrix $\mathbf{C} = [c_{jk}]$, where c_{jk} is the cost of predicting class $\hat{y} = k$ for a data point whose true class is $y = j$. The **classwise expected cost** for predicted class k is given by:

$$C^M_{\mathcal{R}_k} = \mathbb{E}_{p(\mathbf{x},y|\mathbf{x}\in\mathcal{R}_k)}[c_{jk}\mathbb{1}(y = j)] = \sum_{j=1}^{K} c_{jk}\theta_{jk}. \tag{3}$$

The posterior of $C^M_{\mathcal{R}_k}$ is not available in closed form but Monte Carlo samples are straightforward to obtain, by randomly sampling $\theta_{\cdot k} \sim \text{Dirichlet}(\alpha_{\cdot k})$ and computing $C^M_{\mathcal{R}_k}$ deterministically with the sampled $\theta_{\cdot k}$ and the predefined cost matrix $\mathbf{C}$.

Bayesian Estimation of Accuracy Differences Bayesian estimation of group differences allows us to compare the performance between two groups with uncertainty. For example, with prior $\text{Beta}(1, 1)$ and the full test set of CIFAR-100, the posterior distribution of groupwise accuracies of ResNet-110 on superclass "human" and "trees" are $\theta_{g_1} \sim \text{Beta}(280, 203)$ and $\theta_{g_2} \sim \text{Beta}(351, 162)$ respectively. The total amount of labeled data for the two superclasses are 481 and 511. The difference in accuracy between superclass "human" and superclass "trees" is defined as $\Delta = \theta_{g_1} - \theta_{g_2}$. With random samples from the posterior distributions of θ_{g_1} and θ_{g_2}, we can simulate the posterior distribution of Δ and compute its cumulative density in each of three regions $\mu = (P(\Delta < -\epsilon), P(-\epsilon \leq \Delta \leq \epsilon), P(\Delta > \epsilon))$. In Figure 5, when $\epsilon = 0.05$, ResNet-110 is less accurate on the predicted superclass "human" than on "trees" with 96% probability. Similarly with 82% probability, the accuracy of ResNet-110 on "woman" is lower than "man". Although the point estimates of two performance differences have values that are both approximately 10%, the assessment of "human" v.s. "tree" is more certain because more samples are labeled.

Bayesian Assessment: Inferring Statistics of Interest via Monte Carlo Sampling

An additional benefit of the Bayesian framework is that we can draw samples from the posterior to infer other statistics of interest. Here we illustrate this method with two examples.

Bayesian Ranking via Monte Carlo Sampling We can infer the Bayesian ranking of classes in terms of classwise accuracy or expected calibration error (ECE), by drawing samples from the posterior distributions (e.g., see Marshall and Spiegelhalter (1998)). For instance, we can estimate the ranking of classwise accuracy of a model for CIFAR-100, by sampling θ_k's (from their respective posterior Beta densities) for each of the classes and then computing the rank of each class using the sampled accuracy. We run this experiment 10,000 times and then for each class we can empirically estimate the distribution of its ranking. The MPE and 95% credible interval of ranking per predicted class for top 10 and bottom 10 are provided in Figure 6a for CIFAR-100.

Posterior probabilities of the most and least accurate predictions We can estimate the probability that a particular class such as *lizard* is the least accurate predicted class of CIFAR-100 by sampling θ_{k^*}'s (from their respective posterior Beta densities) for each of the classes and then measuring whether θ_{lizard} is the minimum of the sampled values. Running this experiment 10,000 times and then averaging the results, we determine that there is a 68% chance that lizard is the least accurate class predicted by ResNet-110 on CIFAR-100. The posterior probabilities for other classes are provided in Figure 6b, along with results for estimating which class has the highest classwise accuracy.

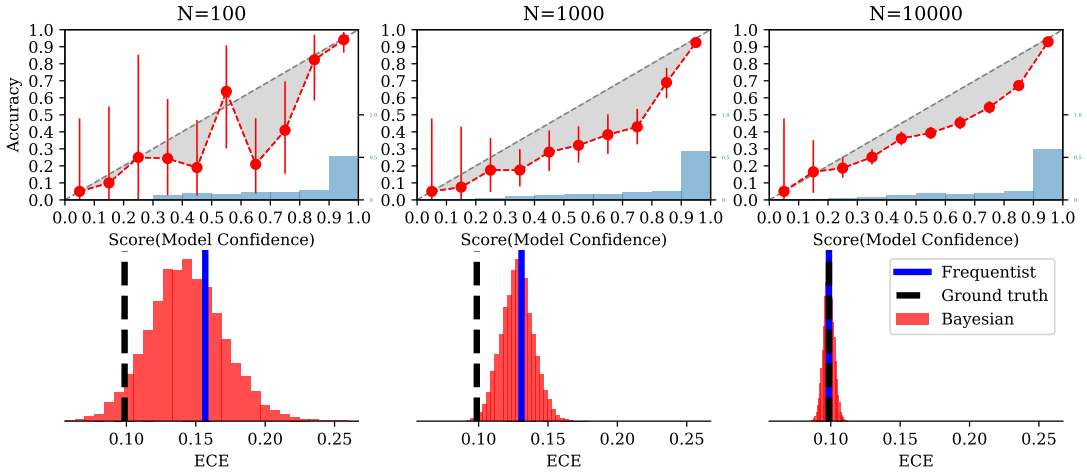


Figure 2: Bayesian reliability diagrams (top) and posterior densities for ECE (bottom) for CIFAR-100 as the amount of data used for estimation increases. Vertical lines in the right plots depict the ground truth ECE (black, evaluated with all available assessment data) and frequentist estimates (red).

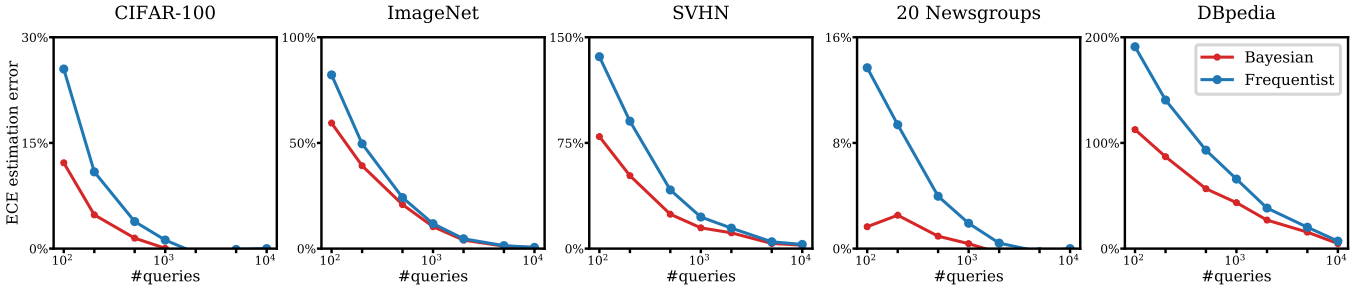


Figure 3: Percentage error in estimating expected calibration error (ECE) as a function of dataset size, for Bayesian (red) and frequentist (blue) estimators, across five datasets.

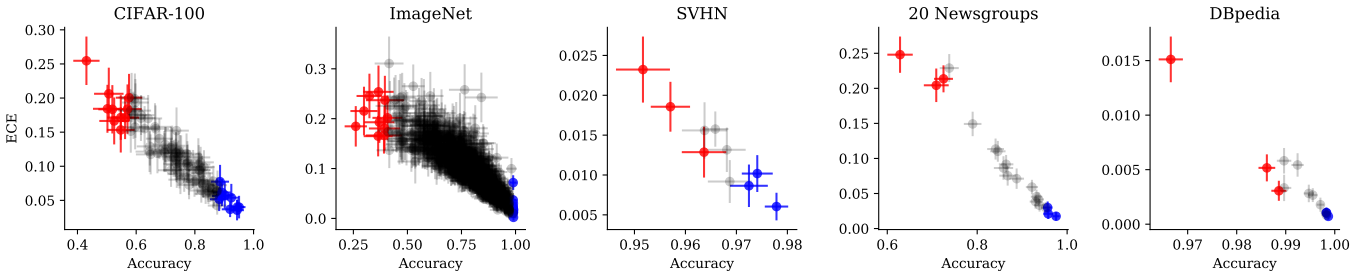


Figure 4: Scatter plots of classwise accuracy and ECE for the five datasets in the main paper. Each marker represents posterior means and 95% credible intervals of posterior accuracy and ECE for each predicted class. Markers in red and blue represent the top-m least and most accurate predicted classes, markers in gray represent the other classes, with $m = 10$ for CIFAR-100 and ImageNet, and $m = 3$ for the other datasets.

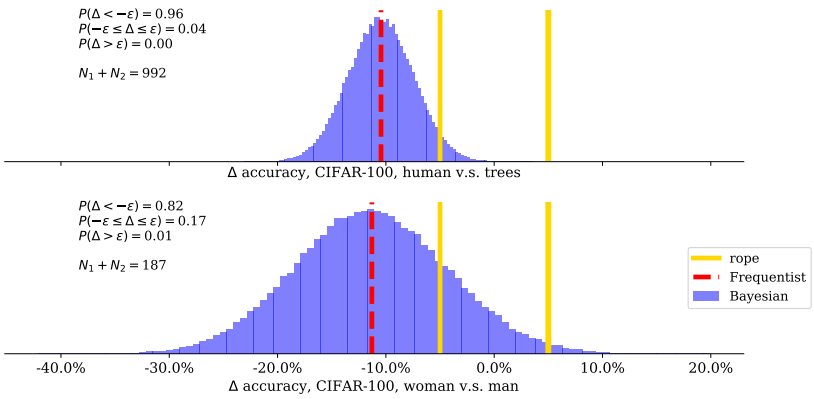


Figure 5: Density plot for the differences of accuracy between two superclasses/classes of CIFAR-100. Region of practical equivalence is [-0.05, 0.05]. Vertical solid lines in gold plots the region of practical equivalence, vertical red dashed line plots the frequentist estimate of accuracy difference. Left: two groups are predicted superclass "human" and "trees" in CIFAR100. Right: two groups are predicted classes "woman" and "man" in CIFAR-100.

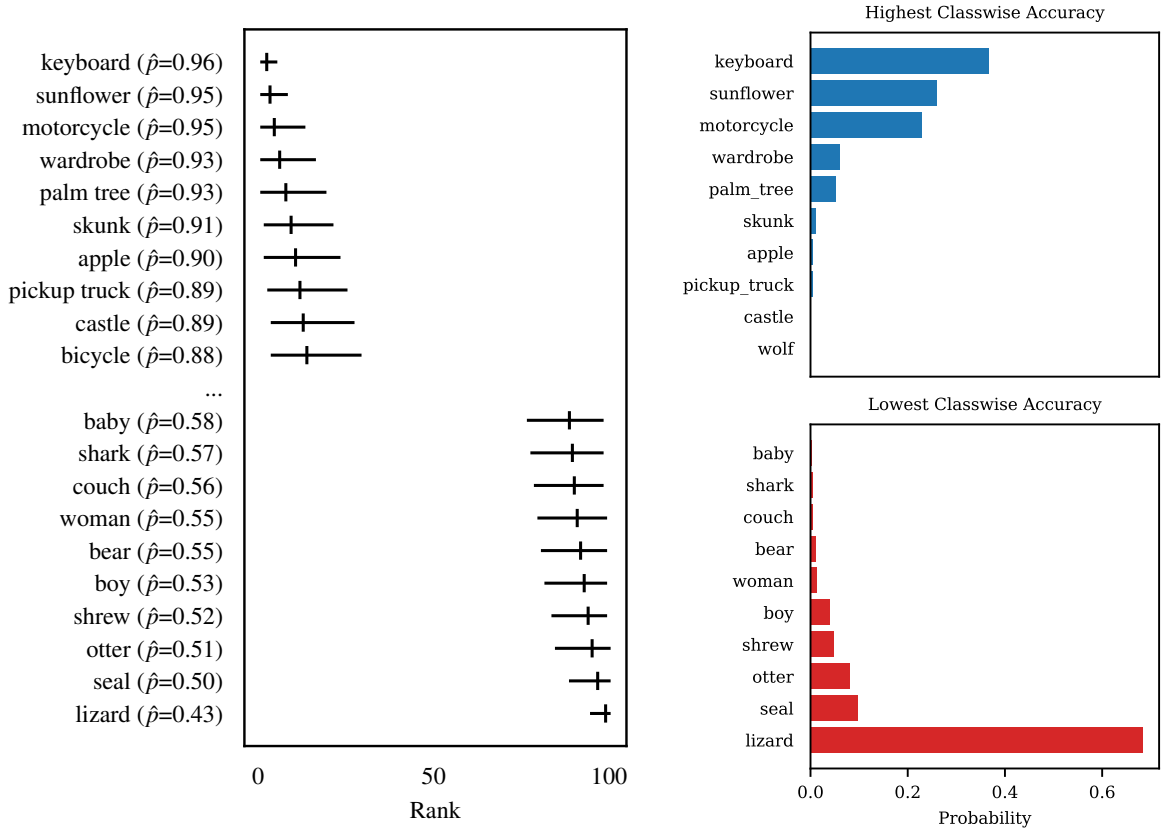


Figure 6: (a) MCMC-based ranking of accuracy across predicted classes for CIFAR-100 (where 1 corresponds to the class with the highest accuracy. (b) Posterior probabilities of the most and least accurate predictions on CIFAR-100. The class with the highest classwise accuracy is somewhat uncertain, while the class with the lowest classwise accuracy is very likely *lizard*.

Supplemental Material for Section 5: Active Assessment

Different Multi-Armed Bandit Algorithms for Best-Arm(s) Identification

Below we provide brief descriptions for Thompson Sampling(TS) and the different variants of multi-armed bandit algorithms for best arm identification and top-m arms identification problems that we investigated in this paper, including Top-Two Thompson Sampling(TTTS) (Russo 2016) and multiple-play Thompson sampling(MP-TS)(Komiyama, Honda, and Nakagawa 2015).

Best Arm Identification

- **Thompson sampling (TS)** is a widely used method for online learning of multi-armed bandit problems (Thompson 1933; Russo et al. 2018). The algorithm samples actions according to the posterior probability that they are optimal. Algorithm 1 describes the sampling process for identifying the least accurate predicted class with TS.
- **Top-two Thompson sampling (TTTS)** is a modified version of TS that is tailored for best arm identification, and has some theoretical advantages over TS. Compared to TS, this algorithm adds a re-sampling process to encourage more exploration.
  Algorithm 2 describes the sampling process for identifying the least accurate predicted class with TTTS. The re-sampling process of TTTS is described in lines 10 to 24. At each step, with probability $1-\beta$ the algorithm selects the class I which has the lowest sampled accuracy; in order to encourage more exploration, with probability β the algorithm re-samples until a different class $J \neq I$ has the lowest sampled accuracy. β is a tuning parameter. When $\beta = 0$, there is no re-sampling in TTTS and it is reduced to TS.

Figure 8 compares TS and TTTS for identifying the least accurate class for CIFAR-100. The results show that two methods are equally efficient across 5 datasets. For TTTS, we set the probability for re-sampling to $\beta = 0.5$ as recommended in (Russo 2016).

Top-m Arms Identification

- **Multiple-play Thompson sampling (MP-TS)** is an extension of TS to multiple-play multi-armed bandit problems and it has a theoretical optimal regret guarantee with binary rewards. Algorithm 3 is the sampling process to identify the least accurate m arms with MP-TS, where m is the number of the best arms to identify. At each step, m classes with the lowest sampled accuracies are selected, as describe in lines 10 to 20. When $m = 1$, MP-TS is equivalent to TS.

In our experiments, we use TS for best arm identification and MP-TS for top-m arms identification, and refer to both of the methods as TS in the main paper for simplicity.

Algorithm 1 Thompson Sampling (TS) Strategy

Input: prior hyperparameters α, β
initialize $n_{k,0} = n_{k,1} = 0$ for $k = 1$ **to** K
repeat
  # Sample accuracy for each predicted class
  for $k = 1$ **to** K **do**
    $\widetilde{\theta}_k \sim \text{Beta}(\alpha + n_{k,0}, \beta + n_{k,1})$
  end for
  # Select a class k with the lowest sampled accuracy
  $\hat{k} = \arg\min_k \widetilde{\theta}_{1:K}$
  # Randomly select an input data point from the $\hat{k}$-th class and compute its predicted label
  $\mathbf{x}_i \sim \mathcal{R}_{\hat{k}}$
  $\hat{y}_i = \arg\max_k p_M(y = k|\mathbf{x}_i)$
  # Update parameters of the $\hat{k}$-th metric
  if $\hat{y}_i = \hat{k}$ **then**
    $n_{\hat{k},0} \leftarrow n_{\hat{k},0} + 1$
  else
    $n_{\hat{k},1} \leftarrow n_{\hat{k},1} + 1$
  end if
until all data labeled

Figure 7: Thompson Sampling (TS) for identifying the least accurate class.

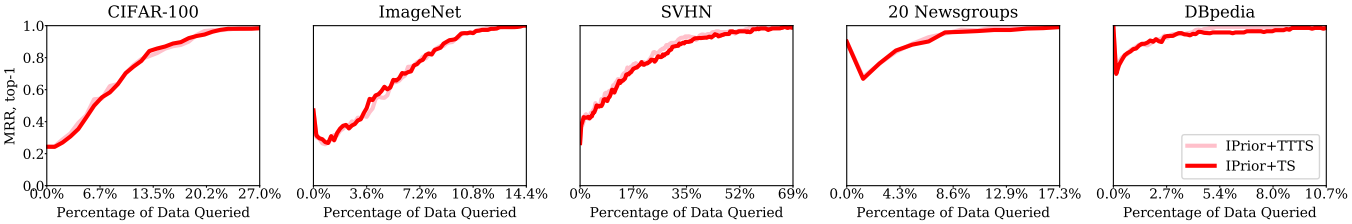


Figure 8: Mean reciprocal rank (MRR) of the class with the estimated lowest classwise accuracy with the strength of the prior set as $\alpha + \beta = 2$, comparing TS and TTTS, across five datasets. The y-axis is the average MRR over 1000 runs for the percentage of queries, relative to the full test set, as indicated on the x-axis.

Algorithm 2 Top Two Thompson Sampling (TTTS) Strategy

Input: prior hyperparameters α, β
initialize $n_{k,0} = n_{k,1} = 0$ for $k = 1$ **to** K
repeat
  # Sample accuracy for each predicted class
  for $k = 1$ **to** K **do**
    $\widetilde{\theta}_k \sim \text{Beta}(\alpha + n_{k,0}, \beta + n_{k,1})$
  end for
  # Select a class k with the lowest sampled accuracy
  $I = \arg\min_k \widetilde{\theta}_{1:K}$
  # Decide whether to re-sample
  $B \sim \text{Bernoulli}(\beta)$
  if $B = 1$ **then**
    # If not re-sample, select I
    $\hat{k} = I$
  else
    # If re-sample, keep sampling until a different arm J is selected
    repeat
      for $k = 1$ **to** K **do**
        $\widetilde{\theta}_k \sim \text{Beta}(\alpha + n_{k,0}, \beta + n_{k,1})$
      end for
      $J = \arg\min_k \widetilde{\theta}_{1:K}$
    until $J \neq I$
    $\hat{k} = J$
  end if
  # Randomly select an input data point from the $\hat{k}$-th class and compute its predicted label
  $\mathbf{x}_i \sim \mathcal{R}_{\hat{k}}$
  $\hat{y}_i = \arg\max_k p_M(y = k|\mathbf{x}_i)$
  # Update parameters of the $\hat{k}$-th metric
  if $\hat{y}_i = \hat{k}$ **then**
    $n_{\hat{k},0} \leftarrow n_{\hat{k},0} + 1$
  else
    $n_{\hat{k},1} \leftarrow n_{\hat{k},1} + 1$
  end if
until all data labeled

Figure 9: Top Two Thompson Sampling (TTTS) for identifying the least accurate class.

Algorithm 3 Multiple-play Thompson sampling (MP-TS) Strategy

Input: prior hyperparameters α, β
initialize $n_{k,0} = n_{k,1} = 0$ for $k = 1$ **to** K
repeat
  # Sample accuracy for each predicted class
  for $k = 1$ **to** K **do**
    $\widetilde{\theta}_k \sim \text{Beta}(\alpha + n_{k,0}, \beta + n_{k,1})$
  end for
  # Select a set of m classes with the lowest sampled accuracies
  $I^* =$ top-m arms ranked by $\widetilde{\theta}_k$.
  for $\hat{k} \in I^*$ **do**
    # Randomly select an input data point from the $\hat{k}$-th class and compute its predicted label
    $\mathbf{x}_i \sim \mathcal{R}_{\hat{k}}$
    $\hat{y}_i = \arg\max_k p_M(y = k|\mathbf{x}_i)$
    # Update parameters of the $\hat{k}$-th metric
    if $\hat{y}_i = \hat{k}$ **then**
      $n_{\hat{k},0} \leftarrow n_{\hat{k},0} + 1$
    else
      $n_{\hat{k},1} \leftarrow n_{\hat{k},1} + 1$
    end if
  end for
until all data labeled

Figure 10: Multiple-play Thompson Sampling (MP-TS) for identifying the least accurate m classes.

Supplemental Material for Section 6: Experiment Settings

Prediction models

For image classification we use ResNet (He et al. 2016) architectures with either 110 layers (CIFAR-100) or 152 layers (SVHN and ImageNet). For ImageNet we use the pretrained model provided by PyTorch, and for CIFAR-100 and SVHN we use the pretrained model checkpoints provided at: https://github.com/bearpaw/pytorch-classification. For text classification tasks we use fine-tuned $\text{BERT}_{\text{BASE}}$ (Devlin et al. 2019) models. Each model was trained on the standard training set used in the literature and assessment was performed on the test sets. Ground truth values for the assessment metrics were computed using the full labeled test set of each dataset.

Evaluation

- **Estimation**: we use RMSE of the estimated $\hat{\theta}$ relative to the true θ^* (as computed from the full test set) to measure the estimation error. For Bayesian methods, $\hat{\theta}$ is the maximum posterior estimation(MPE) of θ's posterior distribution. For frequentist methods, $\hat{\theta}$ is the corresponding point estimation.
  The estimation error of groupwise accuracy and confusion matrix are defined as RMSE $= (\sum_g p_g(\hat{\theta}_g - \theta_g^*)^2)^{\frac{1}{2}}$ and RMSE $= (\sum_k p_k(\sum_j(\hat{\theta}_{jk} - \theta_{jk}^*)^2)^{\frac{1}{2}}$ respectively.
- **Identification**: we compute the mean-reciprocal-rank (MRR) relative to the true top-m ranked groups. The MRR of the top-m classes is defined as $MRR = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{\text{rank}_i}$ where rank_i is the predicted rank of the ith best class. Following standard practice, other classes in the best-m are ignored when computing rank so that $MRR = 1$ if the predicted top-m classes match ground truth. We set $m = 10$ for CIFAR-100 and ImageNet, and $m = 3$ for the other datasets.
- **Comparison**: we compare the results of rope assessment (η, λ) with the ground truth values (η^*, λ^*). The assessment is considered as a success if (1) the direction of difference is correctly identified $\eta = \eta^*$ and (2) the estimation error of cumulative density is sufficiently small $|\lambda - \lambda^*|/\lambda^* < 0.05$.

In all experiments in our paper, unless stated otherwise, we report the aggregated performance averaged over 1000 independent runs.

Reproducibility

We provide code to reproduce our results reported in the paper and in the Appendices. All our datasets and code will be publicly available at: https://github.com/anonymized. The random seeds we used to generate the reported results are provided in the code.

The memory complexity of the non-optimized implementation of Algorithm 1 and 3 is $\mathcal{O}(N + K)$, where N is the number of data points and K is the number of groups. Overall the sampling methods we developed are computationally efficient. For example, for estimating groupwise accuracy of ResNet-110 on CIFAR-100, one run takes less than 10 seconds. All our experiments are conducted on Intel i9-7900X (3.3Ghz, 10 cores) with 32 GB of RAM.

Settings for hyperparameter and priors are discussed in Section "Experimental Settings" in the paper. We set the prior strengths as $\alpha_g + \beta_g = N_0 = 2$ for Beta priors and $\sum \alpha_g = N_0 = 1$ for Dirichlet priors in all experiments, unless otherwise stated, demonstrating the robustness of the settings across a wide variety of contexts. In "Experimental Results: Sensitivity Analysis for Hyperparameters" of this Appendix we provide additional sensitivity analysis.

Cost Matrices

To assess misclassification cost of the models, we experimented with 2 different cost matrices on the CIFAR-100 dataset:

- **Human**: the cost of misclassifying a person (e.g., predicting *tree* when the true class is a *woman*, *boy* etc.) is more expensive than other mistakes.
- **Superclass**: the cost of confusing a class with another superclass (e.g., a *vehicle* with a *fish*) is more expensive than the cost of mistaking labels within the same superclass (e.g., confusing *shark* with *trout*).

We set the cost of expensive mistakes to be 10x the cost of other mistakes. In Figure 11, we plot the two cost matrices.

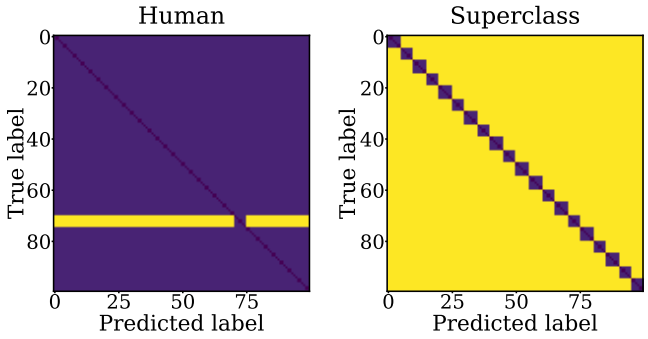


Figure 11: Cost matrices used in our experiments. (left): human, (right): superclass.

Supplemental Material for Section 7: Experimental Results

RMSE of Binwise Accuracy Estimates

In the main paper we report the estimation error results for overall ECE—here we also report results for error per score-bin. Table 1 provides the RMSE results for binwise accuracy estimates. The results demonstrate that both informative priors and active sampling significantly reduce RMSE relative to the baseline for all datasets and all N values.

Table 1: RMSE of **binwise** accuracy estimates obtained with UPrior, IPrior and IPrior+TS across 5 datasets over 1000 independent runs. The strength of priors is 2.

	N/K	N	UPrior (baseline)	IPrior (our work)	IPrior+TS (our work)
CIFAR-100	2	20	23.2	12.8	**11.9**
	5	50	15.7	11.3	**10.0**
	10	100	11.0	9.3	**8.1**
DBpedia	2	20	7.6	2.3	**1.4**
	5	50	3.6	2.5	**1.2**
	10	100	2.4	2.1	**0.9**
20 Newsgroups	2	20	19.6	11.5	**9.0**
	5	50	12.4	9.6	**7.7**
	10	100	8.9	7.6	**6.4**
SVHN	2	20	12.5	4.9	**2.8**
	5	50	6.2	4.5	**2.3**
	10	100	4.4	3.7	**1.9**
ImageNet	2	20	23.0	11.5	**10.6**
	5	50	16.0	10.6	**9.0**
	10	100	11.1	8.8	**7.1**

Identifying the least calibrated class

For identifying the least calibrated classes, in Table 2 we compare the percentage of labeled samples that IPrior and IPrior+TS need to identify the least calibrated top-1 and top-m predicted classes across 5 datasets. Table 2 shows that the improvement in efficiency is particularly significant when the classwise calibration performance has large variance across the classes (as shown in Figure 4), e.g., CIFAR-100, ImageNet and 20 Newsgroups.

Table 2: Percentage of labeled samples needed to identify the least calibrated top-1 and top-m predicted classes.

	ECE, Top 1		ECE, Top m	
Dataset	IPrior	IPrior+TS	IPrior	IPrior+TS
CIFAR-100	88.0	**43.0**	90.0	**59.0**
ImageNet	89.6	**31.0**	90.0	**41.2**
SVHN	58.8	**40.7**	88.4	**77.6**
20 Newsgroups	69.0	**27.9**	90.3	**50.5**
DBpedia	27.9	**8.1**	89.1	**55.6**

Comparisons with Alternative Active Learning Algorithms

There are a variety of other active learning approaches, such as epsilon greedy and Bayesian upper-confidence bound(UCB) methods, that could also be used as alternatives to Thompson sampling.

- Epsilon-greedy: with probability $1 - \epsilon$ the arm currently with the greatest expected reward is selected; with probability ϵ the arm is randomly selected. We set ϵ as 0.1 in our experiments.
- Bayesian upper-confidence bound (UCB): the arm with the greatest upper confidence bound is selected at each step. In our experiments we use the 97.5% quantile, estimated from 10,000 Monte Carlo samples, as the upper confidence bound.

We compare epsilon greedy, Bayesian UCB and Thompson sampling (TS) on the tasks to identify the least accurate and the top-m least accurate predicted classes across five datasets. Figure 12 plots the curves of MRR obtained with three methods as the number of queries increase. We use the uninformative prior with prior strength 2 for all three algorithms. The results show that the MRR curves of Thompson sampling always converge faster than the curves of epsilon greedy and Bayesian UCB, indicating that Thompson sampling is broadly more reliable and more consistent in terms of efficiency for these tasks.

Comparisons Between IPrior+TS and UPrior+TS

In this main paper, we compared UPrior, IPrior and IPrior+TS in experimental results, and left out the results of UPrior+TS due to space limits. In this subsection, we use the comparison between UPrior+TS and IPrior+TS to demonstrate the influence of informative priors when samples are actively labeled for identifying the least accurate top-1 or top-m predicted classes. We set the strength of both the informative prior and the uninformative prior as 2.

The results in Figure 13 illustrate that the informative prior can be helpful when the prior captures the relative ordering of classwise accuracy well (e.g., ImageNet), but less helpful when the difference in classwise accuracy across classes is small and the classwise ordering reflected in the "self-assessment prior" is more likely to be in error (e.g., SVHN, as shown in Figure 4.).

In general, across the different estimation tasks, we found that when using active assessment (TS) informative priors (rather than uninformative priors) generally improved performance and rarely hurt it.

Sensitivity Analysis for Hyperparameters

In Figure 14, we show Bayesian reliability diagrams for five datasets as the strength of the prior increases from 10 to 100. As the strength of the prior increases, it takes more labeled data to overcome the prior belief that the model is calibrated. In Figure 15, we show MRR of the m lowest accurate predicted classes as the strength of the prior increases from 2 to 10 to 100. And in Figure 16, we show MRR of the m least calibrated predicted classes as the strength of the prior increase from 2 to 5 and 10. From these plots, the proposed approach appears to be relatively robust to the prior strength.

Sensitivity to Cost Matrix Values

We also investigated the sensitivity of varying the relative cost of mistakes in our cost experiments. Results are provided in Table 3 and 4. We consistently observe that active assessment with an informative prior performs the best, followed by non-active assessment with an informative prior and finally random sampling.

Table 3: Number of queries required by different methods to achieve a 0.99 mean reciprocal rank(MRR) identifying the class with highest classwise expected cost. A pseudocount of 1 is used in the Dirichlet priors for Bayesian models. The cost type is "Human."

Cost	Top m	UPrior	IPrior	IPrior+TS
1	1	9.6K	9.4K	5.0K
	10	10.0K	10.0K	9.4K
2	1	9.3K	9.3K	4.4K
	10	9.8K	10.0K	8.4K
5	1	9.5K	9.7K	4.5K
	10	9.6K	10.0K	7.9K
10	1	9.3K	9.1K	2.2K
	10	9.6K	9.7K	7.4K

Table 4: Same setup as Table 3. The cost type is "Superclass".

Cost	Top m	UPrior	IPrior	IPrior+TS
1	1	9.9K	10.0K	2.2K
	10	9.8K	9.9K	5.9K
2	1	10.0K	10.0K	2.2K
	10	9.9K	9.9K	5.2K
5	1	9.9K	10.0K	1.8K
	10	9.9K	9.9K	5.3K
10	1	10.0K	9.8K	1.4K
	10	9.9K	9.9K	4.0K

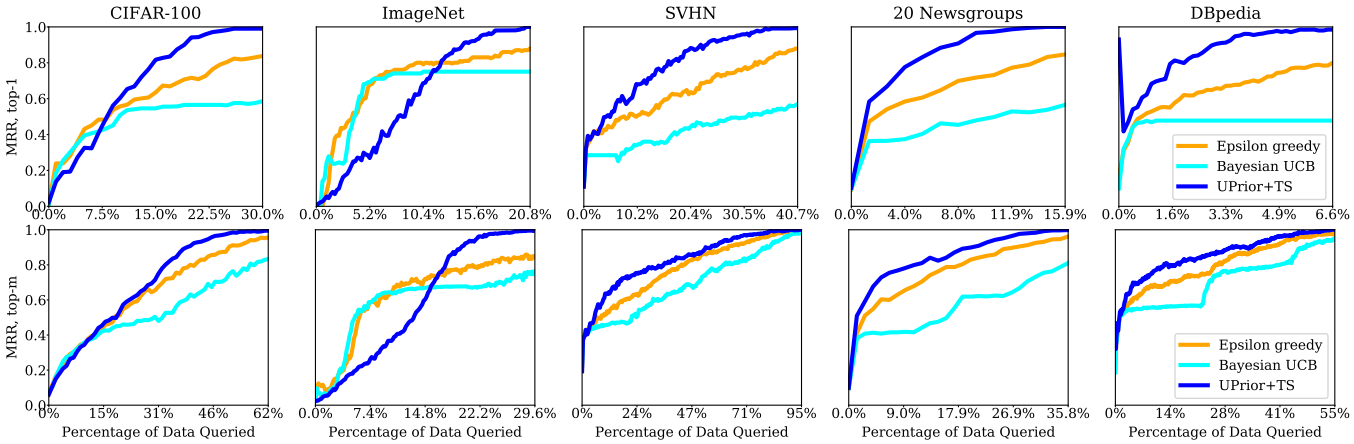


Figure 12: Mean reciprocal rank (MRR) of the classes with the estimated lowest classwise accuracy with the strength of the prior set as 2, comparing Thompson sampling (TS) with epsilon greedy and Bayesian UCB, across five datasets. The y-axis is the average MRR over 1000 runs for the percentage of queries, relative to the full test set, as indicated on the x-axis. In the upper row $m = 1$, and in the lower row $m = 10$ for CIFAR-100 and ImageNet, and $m = 3$ for the other datasets.

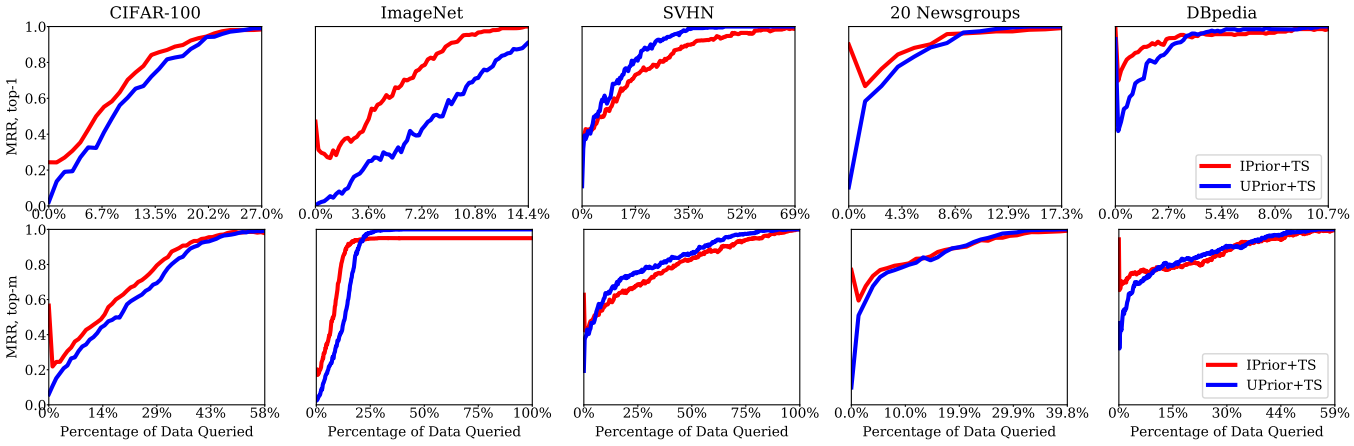
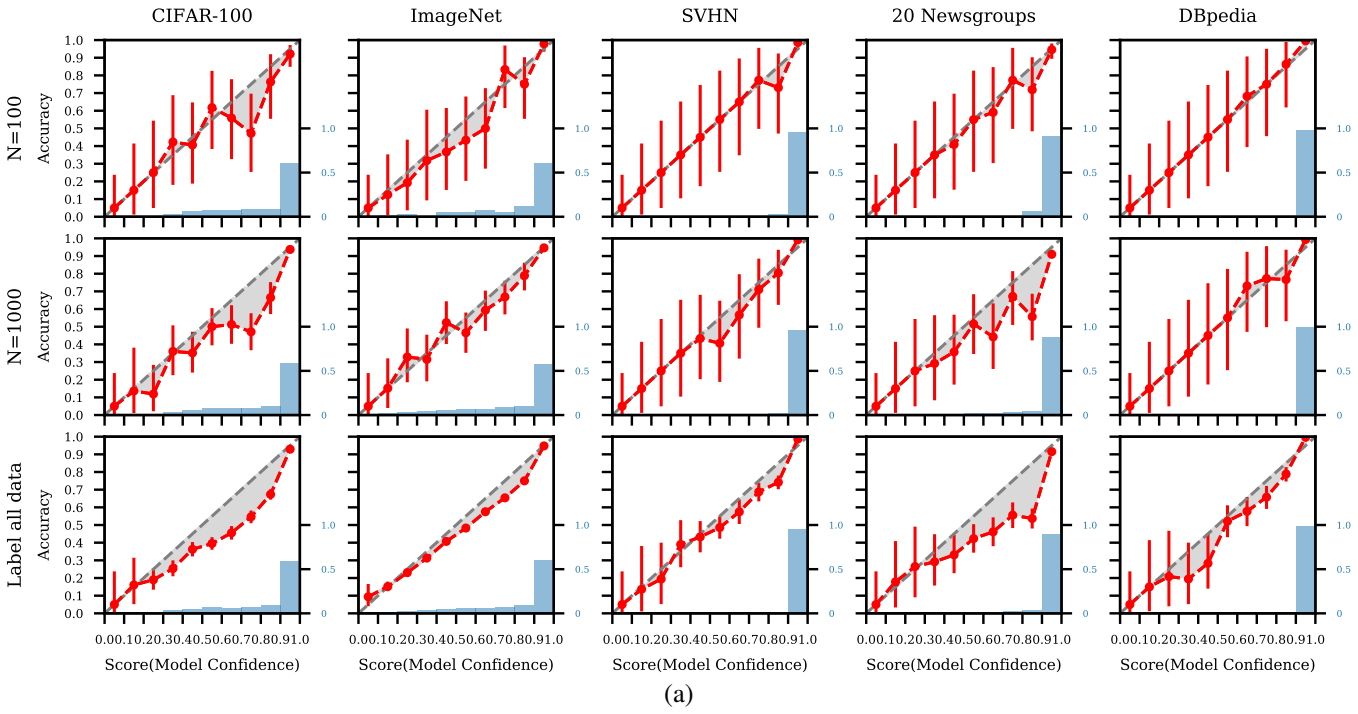
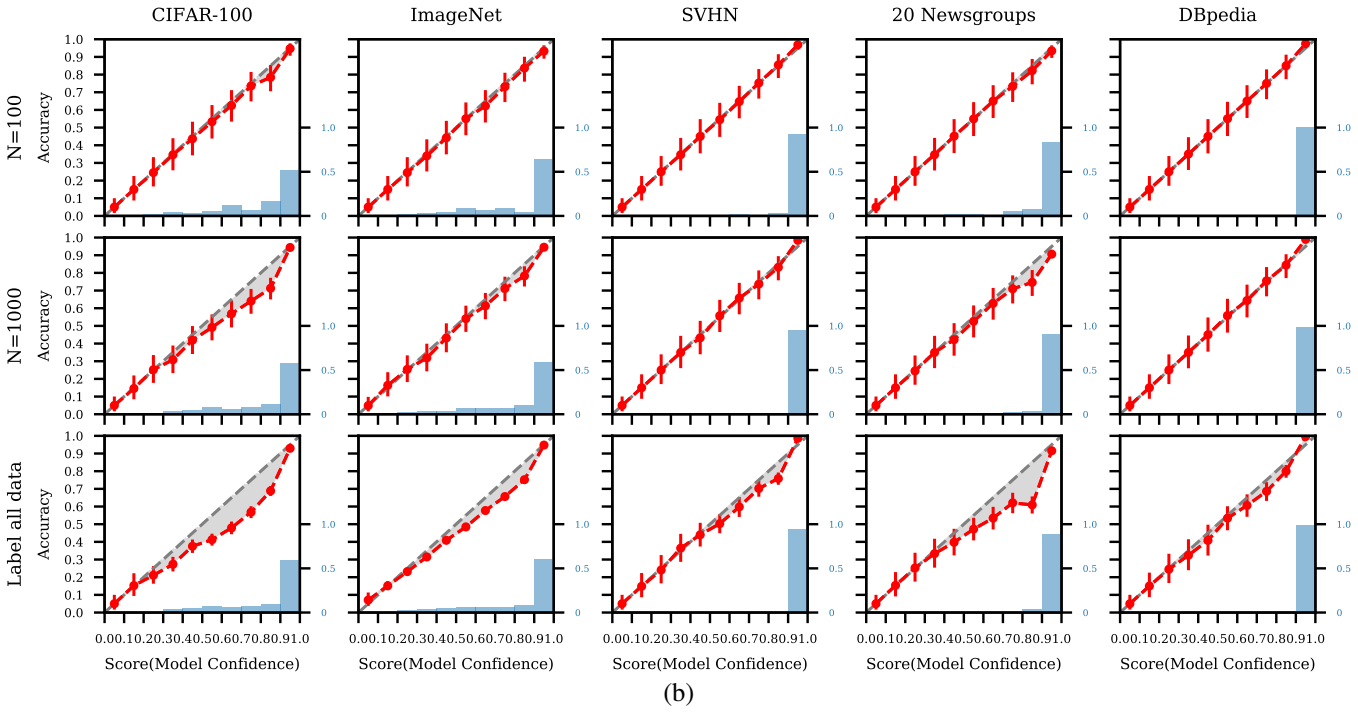


Figure 13: Comparison of the effect of informative (red) and uninformative (blue) priors on identifying the least accurate predicted class with Thompson sampling across 5 datasets. The y-axis is the average MRR over 1000 runs for the percentage of queries, relative to the full test set, as indicated on the x-axis. In the upper row $m = 1$, and in the lower row $m = 10$ for CIFAR-100 and ImageNet, and $m = 3$ for the other datasets.



(a)



(b)

Figure 14: Bayesian reliability diagrams for five datasets (columns) estimated using varying amounts of test data (rows) with prior strength ($\alpha_j + \beta_j$ for each bin) set to be (a) 10 and (b) 100 respectively. The red circles plot the posterior mean for θ_j under our Bayesian approach. Red bars display 95% credible intervals. Shaded gray areas indicate the estimated magnitudes of the calibration errors, relative to the Bayesian estimates. The blue histogram shows the distribution of the scores for N randomly drawn samples.

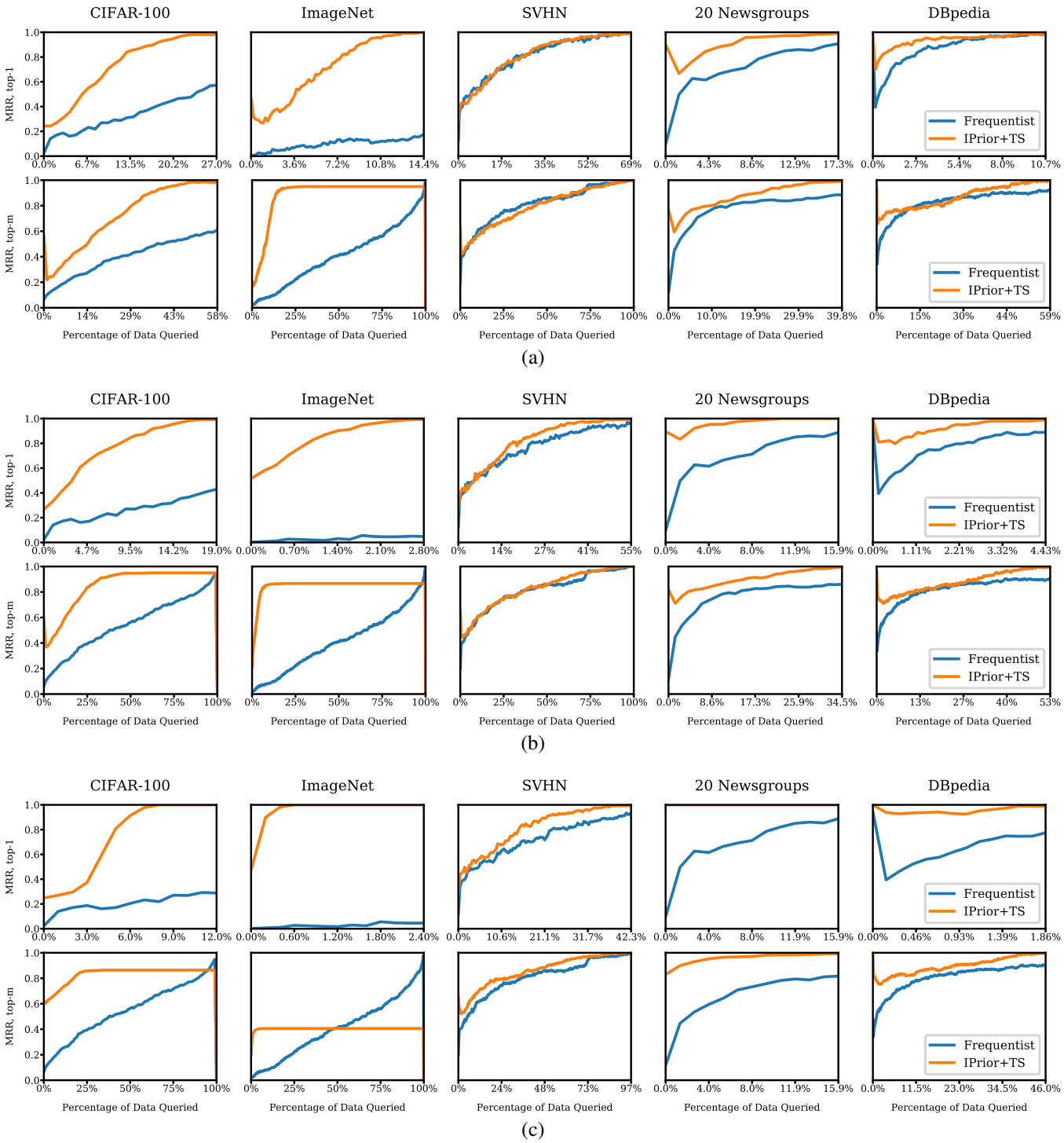
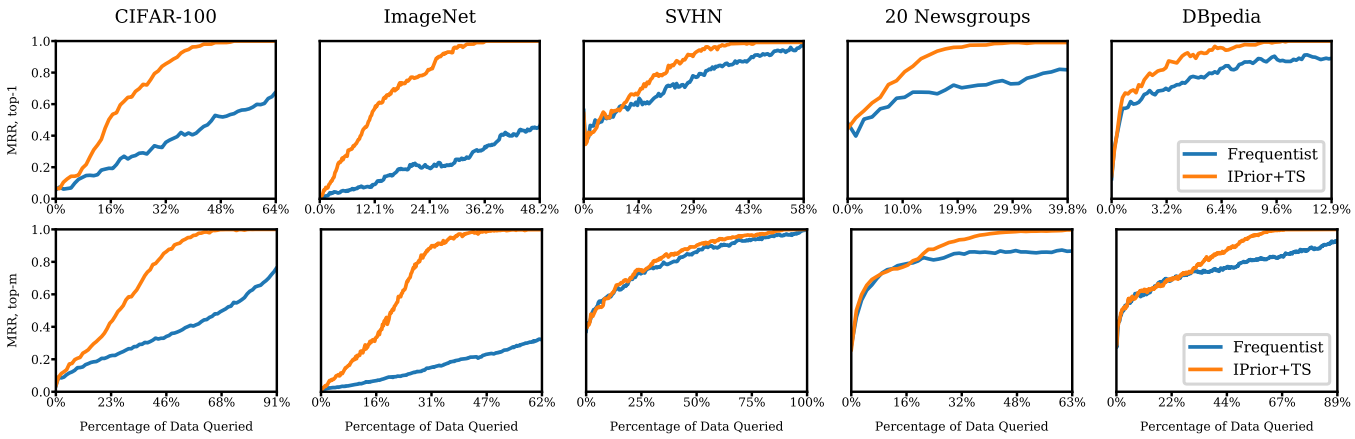
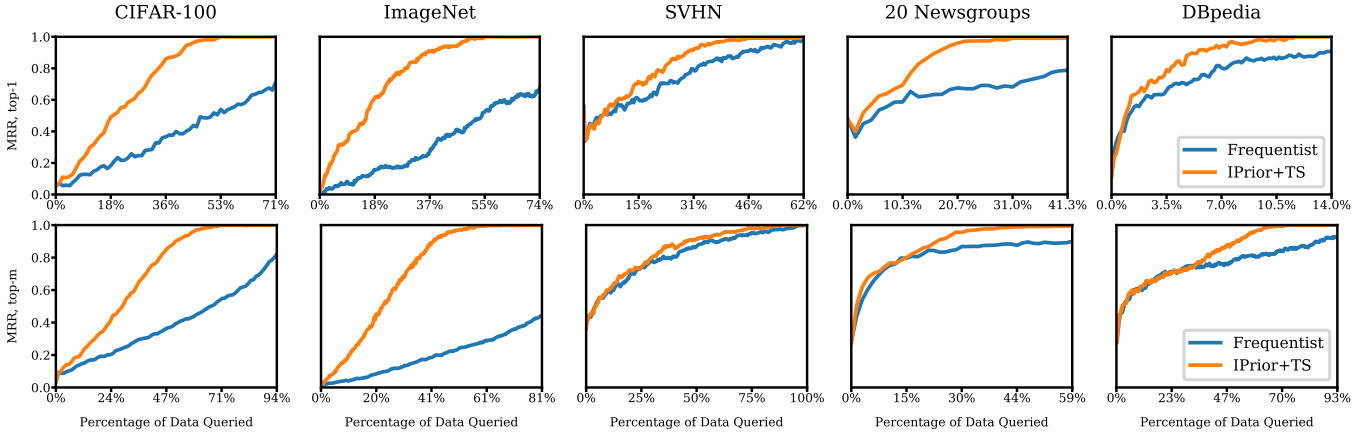


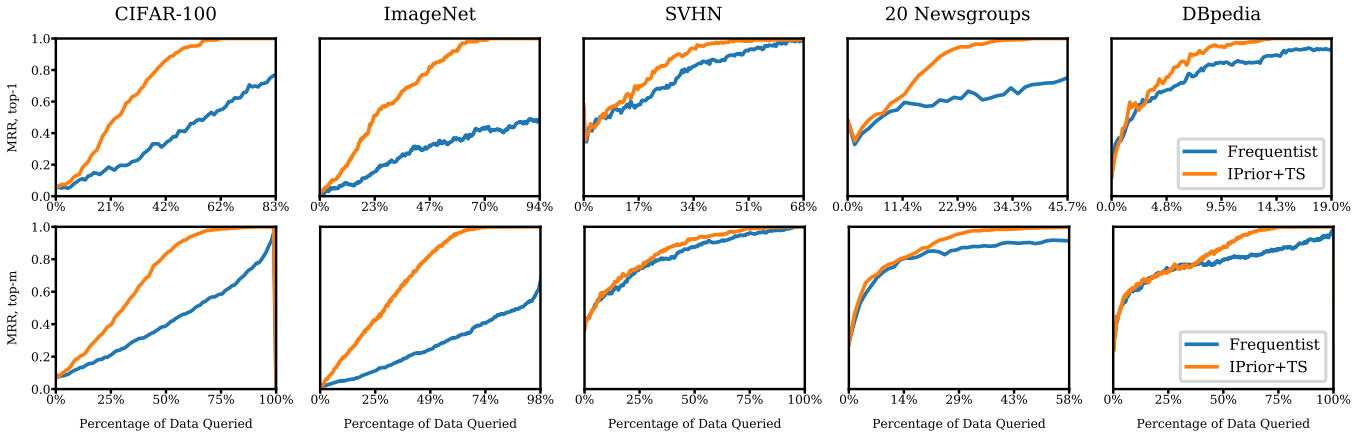
Figure 15: Mean reciprocal rank (MRR) of the m classes with the estimated lowest classwise accuracy as the strength of the prior varies from (a) 2 to (b) 10 and (c) 100, comparing active learning (with Thompson sampling (IPrior+TS)) with no active learning(Frequentist), across five datasets. The y-axis is the average MRR over 1000 runs for the percentage of queries, relative to the full test set, as indicated on the x-axis. For each of (a), (b) and (c), in the upper row $m = 1$, and in the lower row $m = 10$ for CIFAR-100 and ImageNet, and $m = 3$ for the other datasets.



(a)



(b)



(c)

Figure 16: Mean reciprocal rank (MRR) of the m classes with the estimated highest classwise ECE as the strength of the prior varies from (a) 2 to (b) 5 and (c) 10, comparing active learning (with Thompson sampling (IPrior+TS)) with no active learning(Frequentist), across five datasets. The y-axis is the average MRR over 1000 runs for the percentage of queries, relative to the full test set, as indicated on the x-axis. For each of (a), (b) and (c), in the upper row $m = 1$, and in the lower row $m = 10$ for CIFAR-100 and ImageNet, and $m = 3$ for the other datasets.

References

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

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT 2019*, volume 1, 4171–4186.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 1050–1059.

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, 770–778.

Komiyama, J.; Honda, J.; and Nakagawa, H. 2015. Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays. In *International Conference on Machine Learning*, 1152–1161.

Kull, M.; Silva Filho, T.; and Flach, P. 2017. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, 623–631.

Kumar, A.; Liang, P. S.; and Ma, T. 2019. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, 3787–3798.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 6402–6413.

Marshall, E. C.; and Spiegelhalter, D. J. 1998. League tables of in vitro fertilisation clinics: how confident can we be about the rankings. *BMJ* 316: 1701–1704.

Niculescu-Mizil, A.; and Caruana, R. 2005. Predicting good probabilities with supervised learning. In *International Conference on Machine Learning*, 625–632.

Nixon, J.; Dusenberry, M. W.; Zhang, L.; Jerfel, G.; and Tran, D. 2019. Measuring Calibration in Deep Learning. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J. V.; Lakshminarayanan, B.; and Snoek, J. 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, 13969–13980.

Russo, D. 2016. Simple Bayesian algorithms for best arm identification. In *Conference on Learning Theory*, 1417–1418.

Russo, D. J.; Van Roy, B.; Kazerouni, A.; Osband, I.; Wen, Z.; et al. 2018. A tutorial on thompson sampling. *Foundations and Trends in Machine Learning* 11(1): 1–96.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4): 285–294.

Zadrozny, B.; and Elkan, C. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *International Conference on Knowledge Discovery and Data Mining*, 694–699.