Moderating probability distributions for unrepresented uncertainty: Application to sentiment analysis via deep learning

David R. Bickel

Moderating probability distributions for unrepresented uncertainty: Application to sentiment analysis via deep learning

January 13, 2020

David R. Bickel

Ottawa Institute of Systems Biology

Department of Biochemistry, Microbiology and Immunology

Department of Mathematics and Statistics

University of Ottawa

451 Smyth Road

Ottawa, Ontario, K1H 8M5

+01 (613) 562-5800, ext. 8670

dbickel@uottawa.ca

Abstract

The probability distributions that statistical methods use to represent uncertainty fail to capture all of the uncertainty that may be relevant to decision making. A simple way to adjust probability distributions for the uncertainty not represented in their models is to average the distributions with a uniform distribution or another distribution of maximum uncertainty. A decision theoretic framework leads to averaging the distributions by taking the means of the logit transforms of the probabilities. That method does not prevent convergence to the truth, as does taking the means of the probabilities themselves. The mean-logit approach to moderating distributions is applied to natural language processing performed by a deep neural network.

Keywords: big data; data science; deep learning; deep neural network; discounting probability distributions; maximum entropy; unknown loss function

1 Introduction

Statistical models do not incorporate all uncertainty into their probability distributions. As Cox (2001) noted, statistical models only provide lower bounds on uncertainty about a quantity of interest. That is clear in frequentist inference, for each hypothesis test or confidence interval relies on assumptions that remain assumptions even when they pass statistical tests, for the absence of evidence against those assumptions is not evidence for their truth. Bayesian models also underrepresent uncertainty since they could only incorporate all the uncertainty about the models if all reasonable models, their priors, and a hyperprior over the models could be specified with certainty. Even machine learning algorithms fail to capture all uncertainty. For example, neural networks that minimize log loss in effect provide nonparametric maximum likelihood estimates of sampling distributions. They fail to fully quantify the uncertainty in classifications or other predictions, for estimated sampling distributions, as opposed to posterior predictive distributions, neglect the error in the estimates.

Many decision makers are aware that the statistical models they use do not fully represent all uncertainty. They may manage the unrepresented uncertainty by either ignoring it, by hoping it does not negatively impact decisions, or by compensating for it. Such compensation can be informal, without guidance from theory, or formal, with guidance from theory (e.g., Augustin et al., 2014, §4.7).

A simple way to implement the last approach is to adjust reported probabilities for unrepresented uncertainty by combining the reported distribution with a uniform distribution or other distribution of maximum uncertainty, as described in Section 2. A method of distribution combination suitable for that purpose is then proposed in Section 3. Using the framework of the previous two sections, Section 4 specifies the proposed method of moderating a probability distribution to the extent of unrepresented uncertainty and then records its convergence to the truth as the sample size increases. That method is illustrated in Section 5 by using it with a deep neural network to classify movie reviews into two categories: reviews expressing a positive sentiment and those expressing a negative sentiment.

2 Moderating a distribution by combining it with another distribution

Let P_0 denote the *raw distribution*, a probability distribution that may require moderation in order to incorporate unrepresented uncertainty. Let P_1 denote the *moderating distribution*, a probability distribution that would be used for decision making in the extreme case of maximal uncertainty. If uncertainty is measured as entropy and if the domain is finite, then P_1 could be the probability mass function than maximizes the entropy subject to some constraints. In the absence of constraints, that P_1 would be the uniform distribution.

Finally, let P_μ denote a *μ-moderated probability distribution*, where $\mu \in [0,1]$ is the *degree of moderation*. A *method of distribution moderation* is a function that transforms P into P_μ such that P_μ weakly converges to P_0 as $\mu \to 0$ and to P_1 as $\mu \to 1$.

Example 1. In classification problems, machine learning algorithms assign a probability to each of K categories $y(1), \ldots, y(K)$. Let $P_0(y(k) \mid (x,y), x_t)$ denote the reported probability that the category y_t of the independent variable x_t is $y(k)$, where (x,y) is $(x_1,y_1), \ldots, (x_n,y_n)$, a training data set of n pairs of independent variables and their categories, and where each $t = n+1, n+2, \ldots$ is the index of a data pair beyond the training data. Let $P_1(y(k)) = 1/K$ for $k = 1, \ldots, K$.

The μ-moderated probability distribution is $P_\mu(\bullet \mid (x,y), x_t)$, a combination of the probability mass functions $P_0(\bullet \mid (x,y), x_t)$ and $P_1(\bullet)$. Kittler et al. (1998) reviews many methods of averaging classification probabilities over machine learning algorithms. A simple method of combination uses the weighted arithmetic mean (Cooke, 1991), which in this case is

$$P_{\mu,\text{arithmetic}}(y(k) \mid (x,y), x_t) = (1-\mu) P_0(y(k) \mid (x,y), x_t) + \mu P_1(y(k)). \tag{1}$$

Cooke (1991) had considered the weighted arithmetic mean as a way to combine probability distributions elicited from experts. ▲

Example 2. Let $\mathcal{P}$ stand for either $\{P_0, P_1\}$ or the set of mixture distributions of P_0 and P_1:

$$\mathcal{P} = \{(1-w) P_0 + w P_1 : w \in [0,1]\}. \tag{2}$$

With either of those cases of $\mathcal{P}$ as a set of distributions generated by P_0 and P_1, methods of transforming a set of distributions to a single distributions lead to ways to combine P_0 with P_1 in order to moderate P_0.

Several such methods are reviewed by Troffaes (2007) and, in a Bayesian setting, by Bickel (2015). Weighted versions of those methods could be used to transform $\mathcal{P}$ to P_μ by making the weights depend on μ in such a way that P_μ satisfies the above definition of a μ-moderated probability distribution.

The simplest version would simply equate the weight w in equation (2) with μ, yielding $P_\mu = (1-\mu)\,P_0 + \mu\,P_1$. In the special case of supervised classification, that is equivalent to equation (1). An alternative version is described in Section 4. ▲

3 Combining distributions into an adversarial distribution

As noted in Example 2, there are many methods for transforming a set $\mathcal{P}$ of distributions to a single distribution. The transformation method proposed in this section will be applied to that example in Section 4.

The method is based on the following method of combining the probabilities of that $y(k)$ is in some sense the true hypothesis among K mutually exclusive hypotheses $y(1), \ldots, y(K)$. For example, $y(k)$ could represent the hypothesis that a parameter of interest is in an interval of possible parameter values or the hypothesis that the next dependent variable is equal to category $y(k)$. The set of possible probabilities to be combined is denoted by $\mathcal{P}(y(k))$.

Consider future decision makers who must decide whether or not to accept $y(k)$ on the basis of its combined probability. A scale-free, reciprocal invariant distribution of the loss functions of the decision makers leads to

$$\operatorname{logit}^{-1}\left((1-c_k)\operatorname{logit}\overline{P}(y(k)) + c_k \operatorname{logit}\underline{P}(y(k))\right)$$

as the *adversarial probability*, the minimax optimal value of the probability to report to the decision makers (Bickel, 2019), where $\underline{P} = \inf \mathcal{P}(y(k))$, $\overline{P}(y(k)) = \sup \mathcal{P}(y(k))$,

$$\operatorname{logit} P(y) = \log \frac{P(y)}{1-P(y)},$$

and $c_k \in [0,1]$ is the degree of caution toward accepting $y(k)$.

Consider instead the k-independent degree $c(P_0) \in [0,1]$ of caution toward any probability specified by P_0 rather than by P_1, where P_0 and P_1 are any two distributions defined on the same domain. If $\mathcal{P}(y(k)) = \{P_0(y(k)), P_1(y(k))\}$, then the corresponding $c(P_0)$-*adversarial*

probability is

$$\widetilde{P}_{c(P_0)}\left(y\left(k\right)\right) = \text{logit}^{-1}\left(\left(1 - c\left(P_0\right)\right)\text{logit}\, P_0\left(y\left(k\right)\right) + c\left(P_1\right)\text{logit}\, P_1\left(y\left(k\right)\right)\right). \tag{3}$$

That method of combining two probabilities suggests considering $\widetilde{P}_{c(P_0)}$ as the combination of P_0 and P_1, where $\widetilde{P}_{c(P_0)}$ is the function such that $\widetilde{P}_{c(P_0)}\left(y\left(k\right)\right)$ satisfies equation (3) for all $k = 1, \ldots, K$. Since $\widetilde{P}_{c(P_0)}$ need not satisfy $\sum_{k=1}^{K} \widetilde{P}_{c(P_0)}\left(y\left(k\right)\right) = 1$, $\widetilde{P}_{c(P_0)}$ can only be a guess at the combined probability distribution, which is denoted by $P_{c(P_0)}$.

If $\widetilde{P}_{c(P_0)}$ is the initial measure to be updated by a genuine probability distribution $P_{c(P_0)}$ that satisfies intuitively reasonable regularity, locality, transitivity, and weak scaling conditions, then $P_{c(P_0)}$ maximizes entropy in the sense that it minimizes the entropy relative to $\widetilde{P}_{c(P_0)}$, that is,

$$P_{c(P_0)} = \arg \inf_{P': \sum_{k=1}^{K} P'(y(k)) = 1} \sum_{k=1}^{K} P'\left(y\left(k\right)\right) \log\left(\frac{P'\left(y\left(k\right)\right)}{\widetilde{P}_{c(P_0)}\left(y\left(k\right)\right)}\right), \tag{4}$$

according to Csiszár (2008, §4), which summarizes Csiszár (1991). The distribution $P_{c(P_0)}$ satisfying those conditions is called the $c\left(P_0\right)$-*adversarial distribution*.

Lemma 1. *The $c\left(P_0\right)$-adversarial distribution $P_{c(P_0)}$ satisfies*

$$P_{c(P_0)}\left(y\left(k\right)\right) = \frac{\widetilde{P}_{c(P_0)}\left(y\left(k\right)\right)}{\sum_{k'=1}^{K} \widetilde{P}_{c(P_0)}\left(y\left(k'\right)\right)} \tag{5}$$

for all $k = 1, \ldots, K$.

Proof. By definition, $P_{c(P_0)}$ satisfies equation (4), the solution of which is equation (5), as proved using Lagrange multipliers (Jaynes, 2003, §12.3). □

4 The adversarial distribution as the moderated distribution

4.1 The degree of caution as the degree of moderation

This section equates the degree μ to which a raw distribution P_0 is moderated (§2) with the degree $c\left(P_0\right)$ of caution toward P_0 as opposed to P_1 (§3). In short, $\mu = c\left(P_0\right)$. That P_1 would be the distribution for decision making under complete uncertainty, as in Section 2.

Then, for any $c\left(P_0\right)$ between 0 and 1, a $c\left(P_0\right)$-adversarial distribution is a special case of a $c\left(P_0\right)$-moderated distribution. That is formally stated with μ in place of $c\left(P_0\right)$:

Theorem 1. *Consider any $\mu \in [0,1]$. Every μ-adversarial distribution P_μ is a μ-moderated distribution and satisfies*

$$P_\mu\left(y\left(k\right)\right) = \frac{\operatorname{logit}^{-1}\left(\left(1-\mu\right)\operatorname{logit} P_0\left(y\left(k\right)\right) + \mu \operatorname{logit} P_1\left(y\left(k\right)\right)\right)}{\sum_{k'=1}^{K} \operatorname{logit}^{-1}\left(\left(1-\mu\right)\operatorname{logit} P_0\left(y\left(k'\right)\right) + \mu \operatorname{logit} P_1\left(y\left(k'\right)\right)\right)} \tag{6}$$

for all $k = 1, \ldots, K$.

Proof. Plugging μ into the $c\left(P_0\right)$ of equation (3) yields $\widetilde{P}_\mu$. By Lemma 1, every μ-adversarial distribution P_μ satisfies equation (5) with μ substituted for $c\left(P_0\right)$. The substitution of $\widetilde{P}_\mu$ then yields equation (6). Since P_μ weakly converges to P_0 as $\mu \to 0$ and to P_1 as $\mu \to 1$, P_μ is a μ-moderated distribution. □

Example 3. In the notation of Example 1, equation (6) is

$$P_\mu\left(y\left(k\right) \mid \left(x,y\right), x_t\right) = \frac{\operatorname{logit}^{-1}\left(\left(1-\mu\right)\operatorname{logit} P_0\left(y\left(k\right) \mid \left(x,y\right), x_t\right) + \mu \operatorname{logit} P_1\left(y\left(k\right)\right)\right)}{\sum_{k'=1}^{K} \operatorname{logit}^{-1}\left(\left(1-\mu\right)\operatorname{logit} P_0\left(y\left(k'\right) \mid \left(x,y\right), x_t\right) + \mu \operatorname{logit} P_1\left(y\left(k'\right)\right)\right)} \tag{7}$$

rather than equation (1). That expression will be illustrated in Section 5. ▲

4.2 Convergence as the sample size increases

As the sample size n increases, the μ-adversarial distribution converges to the truth, provided that the raw distribution does so, except in degenerate cases such as $\mu = 1$:

Corollary 1. *Suppose there is a category y such that $\lim_{n\to\infty} P_0\left(y\right) = 1$ with probability 1 and $0 < P_1\left(y\right) < 1$. Then $\lim_{n\to\infty} P_\mu\left(y\right) = 1$ with probability 1 for any $\mu < 1$, where P_μ is a μ-adversarial distribution.*

Proof. Assume $0 \leq \mu < 1$. Equation (6) holds by Theorem 1. Thus, since $\lim_{n\to\infty} P_0\left(y\right) = 1$ with probability 1,

$$\begin{aligned}
\lim_{n\to\infty} P_\mu\left(y\right) &\propto \operatorname{logit}^{-1}\left(\left(1-\mu\right)\operatorname{logit} \lim_{n\to\infty} P_0\left(y\right) + \mu \operatorname{logit} P_1\left(y\right)\right) \\
&= \operatorname{logit}^{-1}\left(\left(1-\mu\right) \lim_{n\to\infty} \log \frac{P_0\left(y\right)}{1-P_0\left(y\right)} + \mu \log \frac{P_1\left(y\right)}{1-P_1\left(y\right)}\right) \\
&= \operatorname{logit}^{-1}\left(\left(1-\mu\right) \lim_{n\to\infty} \log \frac{P_0\left(y\right)}{1-P_0\left(y\right)}\right) \\
&= \left(1 + e^{-\left(1-\mu\right)\lim_{n\to\infty} \log \frac{P_0\left(y\right)}{1-P_0\left(y\right)}}\right)^{-1} = \left(1+0\right)^{-1} = 1
\end{aligned}$$

with probability 1. Since $\lim_{n\to\infty} P_0\left(y\right) = 1$ with probability 1 and $\sum_{k=1}^{K} P_0\left(y\left(k\right)\right) = 1$, we have

$\lim_{n\to\infty} P_0\left(y\left(k\right)\right) = 0$ with probability 1 for all $k = 1, \ldots, K$ such that $y\left(k\right) \neq y$. Therefore, for each of those values of k, equation (6) gives

$$\begin{aligned}
\lim_{n\to\infty} P_\mu\left(y\left(k\right)\right) &\propto \operatorname{logit}^{-1}\left((1-\mu)\operatorname{logit}\lim_{n\to\infty} P_0\left(y\left(k\right)\right) + \mu \operatorname{logit} P_1\left(y\left(k\right)\right)\right) \\
&= \operatorname{logit}^{-1}\left((1-\mu)\lim_{n\to\infty}\log\frac{P_0\left(y\left(k\right)\right)}{1-P_0\left(y\left(k\right)\right)} + \mu\log\frac{P_1\left(y\left(k\right)\right)}{1-P_1\left(y\left(k\right)\right)}\right) \\
&= \operatorname{logit}^{-1}\left((1-\mu)\lim_{n\to\infty}\log\frac{P_0\left(y\left(k\right)\right)}{1-P_0\left(y\left(k\right)\right)}\right) = \operatorname{logit}^{-1}\left(-(1-\mu)\lim_{n\to\infty}\log\frac{1-P_0\left(y\left(k\right)\right)}{P_0\left(y\left(k\right)\right)}\right) \\
&= \left(1+e^{+(1-\mu)\lim_{n\to\infty}\log\frac{1-P_0(y(k))}{P_0(y(k))}}\right)^{-1} = 0
\end{aligned}$$

with probability 1. Those two expressions of proportionality together yield the claim since $\sum_{k=1}^{K} P_\mu\left(y\left(k\right)\right) = 1$ according to equation (6). □

That property is highly desirable since it means moderating the raw distribution does not interfere with its convergence as the sample size increases. Intuitively, moderation becomes less necessary as the sample becomes larger. That property is not shared by all moderated distributions; for example, it does not hold for the arithmetic mean of equation (1).

4.3 Decision-theoretic moderation of distributions

For a parameter θ of interest, consider a null hypothesis $y\left(1\right)$ such as $H_0 : \theta = 0$ and a mutually exclusive alternative hypothesis $y\left(2\right)$ such as $H_1 : \theta \neq 0$. The Δ-*discounted posterior probability* of $y\left(1\right)$ is

$$\widetilde{P}^{\star}_{\Delta}\left(y\left(1\right)\right) = \left(1+\left(\frac{P\left(y\left(1\right)\right)}{P\left(y\left(2\right)\right)}\right)^{-\Delta}\right)^{-1} = \left(1+\left(\frac{P\left(y\left(1\right)\right)}{1-P\left(y\left(1\right)\right)}\right)^{-\Delta}\right)^{-1}, \tag{8}$$

which is derived from a decision-theoretic method of moderating posterior distributions, where $\Delta \geq 1$ is the *degree of discounting* (Bickel, 2017, Example 1). The case of no discounting ($\Delta = 1$) then results in $\widetilde{P}^{\star}_{\Delta}\left(y\left(1\right)\right) = P\left(y\left(1\right)\right)$. Analogously, the Δ-*discounted posterior probability* of $y\left(2\right)$ is

$$\widetilde{P}^{\star}_{\Delta}\left(y\left(2\right)\right) = \left(1+\left(\frac{P\left(y\left(2\right)\right)}{P\left(y\left(1\right)\right)}\right)^{-\Delta}\right)^{-1} = \left(1+\left(\frac{P\left(y\left(2\right)\right)}{1-P\left(y\left(2\right)\right)}\right)^{-\Delta}\right)^{-1}. \tag{9}$$

The maximum entropy argument of Section 3 leads to normalizing the discounted probabilities: for $k = 1$ and $k = 2$,

$$P^{\star}_{\Delta}\left(y\left(k\right)\right) = \frac{\widetilde{P}^{\star}_{\Delta}\left(y\left(k\right)\right)}{\widetilde{P}^{\star}_{\Delta}\left(y\left(1\right)\right) + \widetilde{P}^{\star}_{\Delta}\left(y\left(2\right)\right)}. \tag{10}$$

The probability mass function $P^{\star}_{\Delta}$ on $\{y\left(1\right), y\left(2\right)\}$ that satisfies equation (10) is called the Δ-*discounted posterior distribution*. It is a special case of an adversarial distribution:

Corollary 2. *The Δ-discounted posterior distribution is the $(1-1/\Delta)$-adversarial distribution based on the moderating distribution given by $P_1(y(1)) = P_1(y(2)) = 1/2$.*

Proof. By Theorem 1, the $(1-1/\Delta)$-adversarial distribution based on the moderating distribution given by $P_1(y(1)) = P_1(y(2)) = 1/2$ satisfies

$$\begin{aligned}
P_{1-\frac{1}{\Delta}}(y(k)) &= \frac{\operatorname{logit}^{-1}\left(\frac{1}{\Delta}\operatorname{logit} P_0(y(k)) + \left(1-\frac{1}{\Delta}\right)\operatorname{logit}\frac{1}{2}\right)}{\sum_{k'=1}^{2}\operatorname{logit}^{-1}\left(\frac{1}{\Delta}\operatorname{logit} P_0(y(k')) + \left(1-\frac{1}{\Delta}\right)\operatorname{logit}\frac{1}{2}\right)} \\
&\propto \operatorname{logit}^{-1}\left(\frac{1}{\Delta}\log\frac{P(y(k))}{1-P(y(k))} + \left(1-\frac{1}{\Delta}\right)\log\frac{\frac{1}{2}}{1-\frac{1}{2}}\right) \\
&= \operatorname{logit}^{-1}\left(\log\left(\frac{P(y(k))}{1-P(y(k))}\right)^{\frac{1}{\Delta}}\right) \\
&= \left(1+e^{-\log\left(\frac{P(y(k))}{1-P(y(k))}\right)^{\frac{1}{\Delta}}}\right)^{-1} = \left(1+e^{\log\left(\frac{P(y(k))}{1-P(y(k))}\right)^{-\frac{1}{\Delta}}}\right)^{-1} \\
&= \left(1+\left(\frac{P(y(k))}{1-P(y(k))}\right)^{-\frac{1}{\Delta}}\right)^{-1} = \widetilde{P}^{\star}_{\Delta}(y(k))
\end{aligned}$$

for $k = 1, 2$, with the last step following from equations (8) and (9). Since $P_{1-\frac{1}{\Delta}}(y(1)) + P_{1-\frac{1}{\Delta}}(y(2)) = 1$ is required by equation (6), it follows that

$$P_{1-\frac{1}{\Delta}}(y(k)) = \frac{\widetilde{P}^{\star}_{\Delta}(y(k))}{\widetilde{P}^{\star}_{\Delta}(y(1)) + \widetilde{P}^{\star}_{\Delta}(y(2))} = P^{\star}_{\Delta}(y(k))$$

for $k = 1, 2$, with the last step resulting from equation (10). □

That result says that in the cases considered, the degrees of moderation and discounting are related by $\mu = 1 - 1/\Delta$ and $\Delta = 1/(1-\mu)$.

5 Sentiment analysis improved by distribution moderation

5.1 Natural language processing by deep learning

Sentiment analysis is an approach to computational linguistics that automatically extracts opinions from natural language communications between humans. Sentiment analysis has been applied not only to the social sciences and business analytics but also to healthcare (Satapathy et al., 2018). For example, it can be used to infer the quality of care from patients' descriptions of them on social media (Greaves et al., 2013) and can improve the determination of patients' moods from their participation in online networks (Beaunoyer et al., 2017).

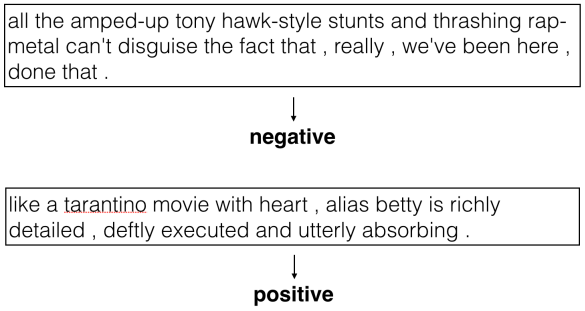

Figure 1: Two string-sentiment pairs randomly selected from the training set.

The hierarchical structure of natural language suggests the use of deep learning via neural networks that have multiple hidden layers (Satapathy et al., 2018, §1.6). Such deep neural networks have outperformed other forms of machine learning for problems in natural language processing (Hasan and Farri, 2019). For instance, Lee et al. (2017) used a deep neural network to detect adverse drug events from social media such as Twitter feeds.

5.2 A sentiment analysis based on a deep neural network

A simple form of sentiment analysis classifies text from a product review as expressing a positive or negative sentiment (Satapathy et al., 2018, §1.5). That problem has been addressed by applying deep neural networks to movie reviews (Radford et al., 2017). The description of that research in Wolfram Research, Inc. (2019b) is summarized in the rest of this subsection.

The sentiment of each movie review is either negative or positive. The training set and non-training set respectively consist of 7462 and 3200 string-sentiment pairs, with each pair consisting of a string of text from a movie review and the sentiment corresponding to the string (Figure 1).

The value of a certain output state of a 27-layer neural network trained on that training set is called a *sentiment score* since it quantifies the sentiment of each string. Replacing the strings in the training set and in the non-training set with their sentiment scores results in a *sentiment training set* and a *sentiment non-training set*, as seen in Figure 2. A naïve Bayes classifier trained on the sentiment training set is surprisingly accurate according to the sentiment non-training set.

5.3 Application of adversarial distributions

The sentiment training set and the sentiment non-training set of Section 5.2 illustrate the method proposed in Section 4. In the notation of Examples (1) and 3, the sentiment scores are the independent variables and the possible sentiments are $y\,(1) = \text{negative}$ and $y\,(2) = \text{positive}$.

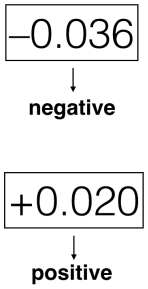

Figure 2: The two score-sentiment pairs from the sentiment training set that correspond to the two pairs displayed in Figure 1.

The pairs in the sentiment non-training set were assigned randomly to a validation set and a test set of equal size. Thus, the sentiment training set, the validation set, and the test set consist of 7462, 1600, and 1600 score-sentiment pairs, respectively. To assess the effect of the sample size on performance, the sentiment training set and validation set were randomly permuted and then reduced to their first 7462ϕ and 1600ϕ pairs for each $\phi \in \{0.01, 0.04, 0.16, 1\}$. After the random permutations, the score-sentiment pair of the ith movie review is (x_i, y_i), where x_i is the sentiment score from the review of a movie and $y_i \in \{\text{negative}, \text{positive}\}$ is the sentiment of the movie for $i = 1, 2, \ldots, 7462 + 3200$.

Let n denote the number of text-sentiment pairs in (x, y), the data set actually used in training a classifier at a value of ϕ. Depending on the classifier, either (x, y) is the sentiment training set of $n = 7462\phi$ text-sentiment pairs, in which case the validation set of 1600ϕ pairs could be used to optimize the degree μ of moderation, or (x, y) is the union of the sentiment training set and the validation set, in which case no data are available for optimizing μ. $\mathcal{T} = \{7462 + 1601, \ldots, 7462 + 3200\}$ is the set of indices of the test set $\left(\left(x_{7462+1601}, y_{7462+1601}\right), \ldots, \left(x_{7462+3200}, y_{7462+3200}\right)\right)$, and $|\mathcal{T}| = 1600$ is the size of the test set, which is not affected by the value of ϕ.

Each classifier considered below is identified by a value of the variable written as clssfr. The test-set mean log-loss and Brier-loss of each classifier are

$$\widehat{\ell}_{\log}(\bullet, \text{clssfr}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} -\log P_{\text{clssfr}}\left(y_t \,|\, (x, y), x_t\right);$$

$$\begin{aligned}\widehat{\ell}_{\text{Brier}}(\bullet, \text{clssfr}) &= \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \sum_{k=1}^{2} \left(P_{\text{clssfr}}\left(y(k) \,|\, (x, y), x_t\right) - \chi_t(k)\right)^2 \\ &= \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left(P_{\text{clssfr}}\left(\text{negative} \,|\, (x, y), x_t\right) - \chi_t(1)\right)^2 + \left(P_{\text{clssfr}}\left(\text{positive} \,|\, (x, y), x_t\right) - \chi_t(2)\right)^2,\end{aligned}$$

where $\chi_t(k) = 1$ if $y(k) = y_t$ and $\chi_t(k) = 0$ if $y(k) \neq y_t$. Then $\chi_t(1)$ indicates whether

y_t = negative, and $\chi_t(2)$ whether y_t = positive. Both loss functions penalize reporting high probabilities of incorrect sentiments.

The mean losses of these classifiers are compared:

1. The data-independent classifier yielding the moderating distribution P_1 that assigns 50% probability to each sentiment: $P_1(\text{negative}) = P_1(\text{positive}) = 1/2$. That uniform classifier is abbreviated as clssfr = "U".

2. Four classifiers using the `Classify` function in Wolfram Research, Inc. (2019a):

   (a) The default logistic regression yields $P_0^{\text{logistic}}(\bullet|(x,y),x_t)$ as the raw distribution $P_0(\bullet|(x,y),x_t)$. The classifier with the training set as (x,y), the data set actually used for training, is denoted by clssfr = "tL", whereas the classifier instead trained on the union of the sentiment training set and validation set as (x,y) is denoted by clssfr = "tvL".

   (b) The default support vector machine yields $P_0^{\text{SVM}}(\bullet|(x,y),x_t)$ as the raw distribution $P_0(\bullet|(x,y),x_t)$. There are two SVM classifiers, named according to the convention used for logistic: clssfr = "tS" and clssfr = "tvS".

3. The eight combined classifiers yielding these versions of the μ-moderated distributions $P_{\mu,\text{arithmetic}}^{\text{logistic}}(\bullet|(x,y),x_t)$, $P_{\mu}^{\text{logistic}}(\bullet|(x,y),x_t)$, $P_{\mu,\text{arithmetic}}^{\text{SVM}}(\bullet|(x,y),x_t)$, and $P_{\mu}^{\text{SVM}}(\bullet|(x,y),x_t)$, formed by combining either $P_0^{\text{logistic}}(\bullet|(x,y),x_t)$ or $P_0^{\text{SVM}}(\bullet|(x,y),x_t)$ with P_1, either according to equation (1) for $P_{\mu,\text{arithmetic}}^{\text{logistic}}$ and $P_{\mu,\text{arithmetic}}^{\text{SVM}}$ ("a" for arithmetic) or according to equation (7) for $P_{\mu}^{\text{logistic}}$ and P_{μ}^{SVM} ("o" for odds-based), and with these degrees of moderation:

   (a) In the simpler case, $\mu = 1/2$ with training logistic or SVM on the union of the sentiment training set and the validation set. In that way, "U" is combined with "tvL" or "tvS". The resulting classifier's name begins with "1/2" and ends with "L" or "S".

   (b) Alternatively, μ, denoted in this case by $\widehat{\mu}$, is fit to minimize the mean log or Brier loss of the validation set after training logistic or SVM on the training data alone. In other words, "U" is combined with "tL" or "tS". The resulting classifier's name begins with "$*$" and ends with "L" or "S".

Each of those 13 classifiers is specified by a value for clssfr according to the above symbols in quotation marks, as follows:

$$\text{clssfr} \in \{\text{"U"}, \text{"tL"}, \text{"tvL"}, \text{"tS"}, \text{"tvS"}, \text{"*oL"}, \text{"}\tfrac{1}{2}\text{oL"}, \text{"*oS"}, \text{"}\tfrac{1}{2}\text{oS"}, \text{"}\tfrac{1}{2}\text{aL"}, \text{"*aL"}, \text{"*aS"}, \text{"}\tfrac{1}{2}\text{aS"}\}. \tag{11}$$

The test-set mean losses of the best-performing classifiers are displayed in Figure 3. When 1% of the sentiment training set and 1% of the validation set is used ($\phi = 1\%$), the $\widehat{\mu}$-moderated SVM distributions ("*oS","*aS") perform better than the SVM distributions ("tS", "tvS"), and the $\widehat{\mu}$-moderated logistic distributions ("*oL","*aL") perform better than the logistic distributions ("tL", "tvL"). When $\phi \in \{4\%, 16\%, 100\%\}$, the same pattern holds for SVM, but the moderation of probabilities does not show a clear advantage for logistic. It thus appears that the original probabilities reported by logistic require no moderation since they adequately reflect the uncertainty about the sentiment. By contrast, the original probabilities reported by SVM, inadequately reflecting the uncertainty about the sentiment, are improved by moderation.

Figure 3 does not indicate a clear advantage of the adversarial distribution ("*oL", "$\frac{1}{2}$oL", "*oS", "$\frac{1}{2}$oS") over the arithmetic mean ("$\frac{1}{2}$aL", "*aL", "*aS", "$\frac{1}{2}$aS"). That suggests that the sentiment training set is too small for Corollary 1 to be relevant.

Acknowledgments

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN/356018-2009).

References

Augustin, T., Coolen, F., de Cooman, G., Troffaes, M. (Eds.), 2014. Introduction to Imprecise Probabilities. Wiley Series in Probability and Statistics. Wiley.

Beaunoyer, E., Arsenault, M., Lomanowska, A. M., Guitton, M. J., 2017. Understanding online health information: Evaluation, tools, and strategies. Patient Education and Counseling 100 (2), 183 – 189.

Bickel, D. R., 2015. Inference after checking multiple Bayesian models for data conflict and applications to mitigating the influence of rejected priors. International Journal of Approximate Reasoning 66, 53–72.

Bickel, D. R., 2017. Departing from Bayesian inference toward minimaxity to the extent that the posterior distribution is unreliable, working paper, HAL-01673783.
URL https://hal.archives-ouvertes.fr/hal-01673783

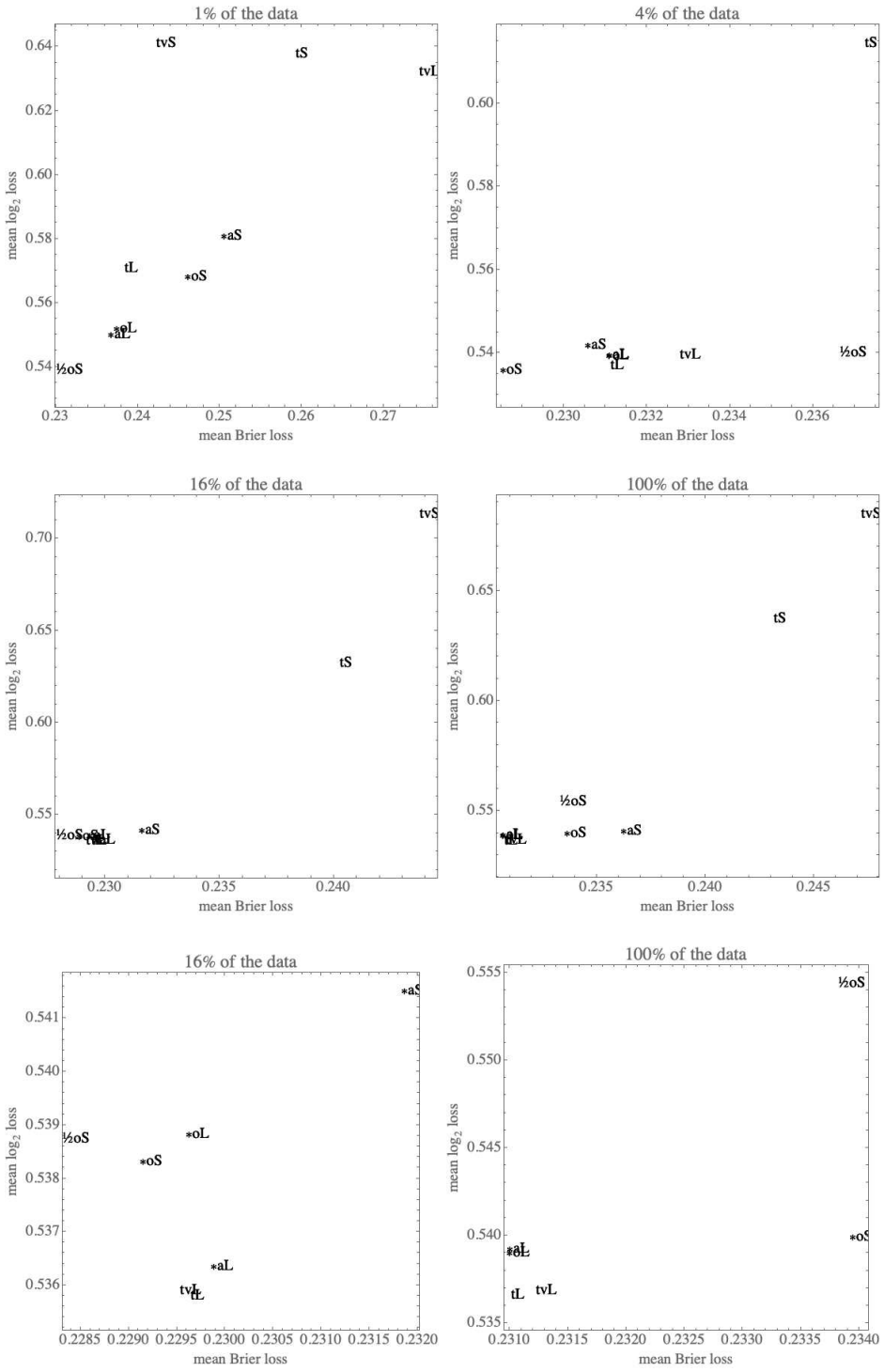

Figure 3: Test-set mean losses $\widehat{\ell}_{\log}(\bullet, \text{clssfr})$ and $\widehat{\ell}_{\text{Brier}}(\bullet, \text{clssfr})$ for sentiment classification, where each classifier is represented by a prefix and then a suffix, as follows. The suffixes indicate either the logistic regression classifier ("L") or the support vector machine classifier ("S"). The prefixes refer to training the classifier on the sentiment training set alone ("t"), training the classifier on the sentiment training set and the validation set ("tv"), or averaging the classifier with the uniform distribution using either 50% moderation and the validation set for training ("1/2") or a degree of moderation fitted to the validation set ("∗") and using either the arithmetic mean ("a") or the odds-based mean ("o"). (Each displayed classifier appears as a value of clssfr in expression (11).) Each of the first four plots corresponds to a different fraction ϕ of the training and validation sets, and the last two plots zoom in on the very best performers when needed to distinguish them from each other.

Bickel, D. R., 2019. Reporting Bayes factors or probabilities to decision makers of unknown loss functions. Communications in Statistics - Theory and Methods 48, 2163–2174.

Cooke, R. M., 1991. Experts in Uncertainty: Opinion and Subjective Probability in Science. Oxford University Press.

Cox, D. R., 2001. Comment on 'Statistical modeling: The two cultures'. Statistical Science 16 (3), 216–218.

Csiszár, I., 1991. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. Ann. Stat. 19, 2032–2066.

Csiszár, I., 2008. Axiomatic characterizations of information measures. Entropy 10, 261–273.

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., Donaldson, L., 2013. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. BMJ Quality & Safety 22 (3), 251–255.

Hasan, S. A., Farri, O., 2019. Clinical Natural Language Processing with Deep Learning. Springer International Publishing, Cham, pp. 147–171.

Jaynes, E., 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge.

Kittler, J., Hatef, M., Duin, R. P. W., Matas, J., March 1998. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (3), 226–239.

Lee, K., Qadir, A., Hasan, S. A., Datla, V., Prakash, A., Liu, J., Farri, O., 2017. Adverse drug event detection in tweets with semi-supervised convolutional neural networks. In: Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp. 705–714.
URL https://doi.org/10.1145/3038912.3052671

Radford, A., Józefowicz, R., Sutskever, I., 2017. Learning to generate reviews and discovering sentiment. CoRR abs/1704.01444.
URL http://arxiv.org/abs/1704.01444

Satapathy, R., Cambria, E., Hussain, A., 2018. Sentiment Analysis in the Bio-Medical Domain: Techniques, Tools, and Applications. Socio-Affective Computing. Springer International Publish-

ing, New York.

URL https://books.google.ca/books?id=gbZIDwAAQBAJ

Troffaes, M. C. M., 2007. Decision making under uncertainty using imprecise probabilities. International Journal of Approximate Reasoning 45 (1), 17–29.

Wolfram Research, Inc., 2019a. Mathematica, version 12.0.0.0.

URL https://www.wolfram.com

Wolfram Research, Inc., 2019b. Sentiment Language Model Trained on Amazon Product Review Data.

URL http://bit.ly/30em3aN