Calibrated regressors for semi-supervised learning on sequences

Alan Amin
For internal use by Jura Bio (do not share)

August 2023

1 Introduction

A few weeks ago Pierre Glaser from Arthur Gretton's group reached out to me to discuss building a test for calibration using the techniques of my KSD paper. In our discussions we decided that calibration of the posterior is a critical consideration during design. I connected him with Steffan in the Marks lab and they are working on a project. Here I ask how calibration of a regressor like T cell match affects calibration of the posterior.

The motivation for this exposition is that calibration is usually motivated by an argument that is not immediately applicable to design: in the case of safety, the idea is that one makes an observation, makes a prediction from that observation, and then must make a decision based on the prediction alone – the prediction must be accurate to guarantee the safety of this decision. But why do we need a calibrated predictor if we're doing design, and what can calibration tell us about the accuracy of our designs? In these notes I also address how to measure calibration for counts data, review tips to build a calibrated regressor, and also how to measure calibration when one only has hits.

2 Bias - calibration decomposition

Say we want to design sequences X conditional on a property y, given a known prior $p(X)$. We start by learning from experimental measurements of sequences to get a predictor $\mu_\theta(X) = p_\theta(y|X)$. Next we want to use Bayes' Rule to form a posterior for X:

$$p_\theta(X|y) = \frac{p_\theta(y|X)p(X)}{\int dp(X')p_\theta(y|X')}.$$

What is the error of this posterior in terms of μ_θ? In these notes I will argue that the calibration of μ_θ is strongly related to the calibration of this posterior, and thus important for the design question.

We can decompose the error into two terms:

$$\begin{aligned} p(X|y) - p_\theta(X|y) = &(p(X|y) - \hat{p}_\mu(X|y)) \\ &+ (\hat{p}_\mu(X|y) - p_\theta(X|y)) \end{aligned} \tag{1}$$

where

$$\hat{p}_\mu(X|y) = \frac{p(y|\mu_\theta(X))p(X)}{\int dp(X')p(y|\mu_\theta(X'))} = \frac{p(y|\mu_\theta(X))p(X)}{p(y)} = p(X|\mu(X))p(\mu(X)|y)$$

is the posterior from the exact joint distribution, conditioning y only on the prediction of the model. The first term is the bias caused by the capacity of the predictor μ_θ – it is how much information that is relevant for predicting y in X is lost by applying the function $\mu_\theta(X)$ and forgetting X. The error can be bounded by

$$E_y\mathrm{KL}(p(X|y)||\hat{p}_\mu(X|y)) = E\mathrm{MI}(y, X|\mu_\theta(X))$$

which is upper bounded by the entropies of X and y. For large models, this error is usually negligible [14]. Thus, I argue that it can be ignored.

The second term is the error from miscalibration, which will be our focus. First note that this term can be written as

$$p(X|\mu(X))\left(p(\mu_\theta(X)|y) - p_\theta(\mu_\theta(X)|y)\right)$$

and this its magnitude is determined by the bracketed term. If our goal is to measure a discrepency then we have significantly simplified the problem by moving from looking at distributions of X to looking at distributions of $\mu_\theta(X)$: X is a sequence, which is a complex, high dimensional object, $\mu_\theta(X)$ is a distribution of y with usually only one or two parameters, for example, $\mu_\theta(X)$ may simply be the mean of a Poisson distribution. Now the error can be decomposed again as

$$p(X|\mu_\theta(X))\left(\frac{1}{p(y)}\left(\hat{p}_\mu(y, \mu_\theta(X)) - p_\theta(y, \mu_\theta(X))\right) + p_\theta(y, \mu_\theta(X))\left(\frac{p(y)}{p_\theta(y)} - 1\right)\right).$$

Calling the measure on the space of sequences $\epsilon(y) = \hat{p}_\mu(y, \mu_\theta(X)) - p_\theta(y, \mu_\theta(X))$, the error can be written

$$p(X|\mu_\theta(X))\left(\frac{1}{p(y)}\epsilon(y) + p_\theta(y, \mu_\theta(X))\left(\frac{\epsilon(y)(S)}{p(y) - \epsilon(y)(S)}\right)\right)$$

where S is the set of all sequences. Clearly to send this error to 0 we must ensure that $\epsilon(y)$ is small.

3 Measuring calibration

Let k be a kernel on the domain of y and c a kernel on the space of distributions on the domain of y. We will try to upper bound the following average error, which is related to the calibration of the posterior:

$$\sup_{\|f\|_k, \|g\|_c \leq 1} \frac{1}{E_{p(y)}f(y)} |E_{p(y)}f(y)\left(E_{\hat{p}_\mu(X|y)}g(X) - E_{p_\theta(X|y)}g(X)\right)|. \tag{2}$$

This can be interpreted as picking a y from your data weighted by f (naturally we pick y within the dynamic range of our experiments, but we care more about larger values, say), designing sequences according to your posterior, then measuring $g(X)$[1]. I'll also note again here that we are interested in measuring the performance of μ_θ, and then working out the above described error of the posterior if we had a perfect $p(X)$, which we often don't. The above error can be bounded by the calibration of the posterior $p_\theta(X|y)$ directly if we were confident in our prior [6]. Instead we will attempt to bound it by the calibration of μ_θ.

Now we define the calibration of the predictions $\mu_\theta(X)$; We will afterwards try to show that if the regressor is nearly calibrated, then the error described above is small. Calibration is usually measured using ECE or MCE which are only immediately defined for when y belongs to a finite set [20]. A more flexible approach is to measure the error is to use MMD [21, 6]

$$\begin{aligned}\mathrm{MMD}(\hat{p}_\mu(y,X), p_\theta(y,X)) &= \sup_{\|f\|_{k\otimes c}\leq 1} |E_{\hat{p}_\mu(y,X)} f(y,\mu_\theta(X)) - E_{p_\theta(y,X)} f(y,\mu_\theta(X))| \\ &= \sup_{\|f\|_{k\otimes c}\leq 1} |E_{p(X)} \left(E_{p(y,\mu_\theta(X))} f(y,\mu_\theta(X)) - E_{\mu_\theta(X)} f(y,\mu_\theta(X))\right)| \\ &= \sup_{\|f\|_{k\otimes c}\leq 1} |\sum_y \int d\epsilon(y)(\mu) f(y,\mu)|.\end{aligned} \tag{3}$$

This MMD is turned into a p-value for a two sample test using one of the methods in [8].

3.1 Inequalities to bound posterior error

Assume $1 \in \mathcal{H}_c$ and $\|1\|_c = 1$(simply redefine $c'(\mu,\nu) = c(\mu,\nu) + 1$), so that $\mathcal{H}_k \otimes 1 \subset \mathcal{H}_k \otimes \mathcal{H}_c$. Then

$$\mathrm{MMD} \geq \sup_{\|f\otimes 1\|_{k\otimes c}\leq 1} |\sum_y f(y)\epsilon(y)(S)|.$$

With this, if k is bounded by 1, the error eqn 2 can be bounded by

$$\begin{aligned}&\frac{1}{E_{p(y)}f(y)}\left(\mathrm{MMD} + \sup_f \left|\sum_y f(y)\frac{\epsilon(y)(S)}{1-p(y)\epsilon(y)(S)}\right|\right) \\ &\leq \frac{1}{E_{p(y)}f(y)}\left(\mathrm{MMD} + \sup_f \left|\sum_y f(y)\epsilon(y)(S)\right|\right. \\ &\qquad \left. + \sup_f \left|\sum_y f(y)\epsilon(y)(S)^2 \frac{p(y)}{1-p(y)\epsilon(y)(S)}\right|\right) \\ &\leq \frac{1}{E_{p(y)}f(y)}\left(2\mathrm{MMD} + \frac{\|\epsilon(y)(S)\|_2^2}{1-\|\epsilon(y)(S)\|_2}\right).\end{aligned}$$

[1] Notice that we could measure the discrepancy between $p_\theta(X|y)$ and the true posterior $p(X|y)$ with a very similar expression. Our motivation for looking at this expression is that sequences are complex and high dimensional, while $\mu(X)$ is simple.

We can bound

$$\|\epsilon(y)(S)\|_2 = \sup_{\|f\otimes 1\|_{k\otimes c}\leq 1} |\sum_y f(y)\epsilon(y)(S)|$$

by picking k to be the identity kernel (perhaps we should use two measurements of calibration: one using a regular kernel and one with the identity kernel; but will the identity kernel measurement take a long time to converge?).

Note the following natural interpretation of this upper bound: to get small error around y's prioritized by f, we must make sure our MMD is small compared to how commonly we see those prioritized y's $E_{p(y)}f(y)$.

4 Building calibrated regressors

From a statistical learning theory point of view, calibration is another objective for the regressor; to get a more calibrated regressor, calibration must be phrased as an empirical risk which must be minimized, and whose minimization must be traded off with model fit or other desiderata [10, 4, 5]; these methods unfortunately can be complicated to implement and can harm accuracy. Another branch of work takes a non-probabilistic model and makes it calibrated after the fact [15, 13] (conformal inference fits in this category); however these methods do not immediately fit into a principled Bayesian inference framework – their application to modelling and design is non-obvious [3, 16]. There are also other heuristics associated to calibration, such as model size [7, 17].

Another branch of work reasons that since Bayesian inference is calibrated [2], building a calibrated regressor with a neural network hinges on performing accurate Bayesian inference with a good prior [22]. This branch of work claims that heuristics that lead to more accurate Bayesian inference [12, 11] result in both more accurate and calibrated predictions! This reasoning is also why it is claimed that settings in which the posterior is easier to capture, such as when a fully connected layer is replaced with a GP[23, 18], or better modelling noise [9, 19], result in more accurate and calibrated predictions. These later techniques of ensembling, SWA, and learning a more flexible emission distribution are the lowest hanging fruit in my opinion.

5 Practical measures of calibration for counts data

The kernel we will use is $\tilde{k} = c\otimes k$ where k is a kernel on the domain of y and c is a kernel on the space of measures on the domain of y most naturally chosen to be $c(\mu,\nu) = \exp(-\frac{1}{2\tau^2}\|\mu-\nu\|_k^2)$ which is universal under some conditions [1]([21] used the parameterization of the model to define c, but then MMDs are not comparable, [6] used a different score-based kernel; another choice for c is an embedding kernel on $\mu\mapsto(E\mu, \mathrm{Var}(\mu))$ or some other variation). Then the empirical MMD for data $((X_i, y_i))_i$ is

$$\sum_i \sum_j c(\mu(X_i), \mu(X_j))(\mu(X_i) - y_i|\mu(X_j) - y_j)_k. \tag{4}$$

To calculate this efficiently, we want a kernel such that the kernel embedding of a measure is easy to calculate, or we could approximate

$$(\mu|y)_k = E_{y' \sim \mu} k(y, y')$$

empirically. An empirical witness function on observations (X_i, y_i) can also be similarly defined. In our case, μ could be a Poisson or negative binomial (or sums of these distributions, in the case of an ensemble or SWA).

For now, a kernel universal on the set of sums of Poisson and negative binomial distributions (negative binomial variance is always greater than its mean – although this is not true if zero-inflation is included, so another feature may need to be included), use

$$c(\mu, \nu) = \exp(-\frac{1}{2\tau^2}\left((E\mu - E\nu)^2 + (\text{std}(\mu) - \text{std}(\nu))^2\right))$$

and use $k(y, y') = \exp\left(-\frac{1}{2\sigma^2}(\tilde{y} - \tilde{y}')^2\right)$ where $\tilde{y} = \log(y + 1)$; empirically estimate $(\mu|y)_k$.

6 Measuring calibration with only hits

In this section we consider the case when we've only observed hits. In this case $y \in \{0, 1\}$ and we want to make sure $\epsilon(0)$ and $\epsilon(1)$ are small. ECE and MCE are the classical tool for measuring this error [20] and can be written as IPMs with $\|f\|_\infty < 1$ and $\|\|f\|_{\infty,y}\|_{1,X} < 1$ [21]. In particular,

$$ECE = \|\epsilon(0)\|_{\text{TV}} + \|\epsilon(1)\|_{\text{TV}}$$

Calibration is usually shown with a calibration curve like figure 3 of [9]. The ECE is the average deviation of teh calibration curve and the MCE is the maximum deviation.

Say we know the total number of sequences N and we have n hits. Elsewhere I showed that if we sample N $X'_1, \ldots, X'_N \sim p(X)$ and label them as non-hits, then we are sampling from a distribution $\tilde{p}(X, y)$ such that

$$\tilde{p}(X, y) = \frac{1}{1 + p(y = 1)} p(X, y) + \frac{p(y = 1)}{1 + p(y = 1)} p(X|y = 1) \otimes \delta_{y=0}.$$

In this case, p_θ is calibrated if and only if the predictor $\frac{\mu_\theta}{1+\mu_\theta}$ is calibrated on the above data. The calibration curve in this case is exactly the full calibration curve transformed by $x \mapsto \frac{x}{1+x}$ on $[0, 1]$. This observation tells us that we can calculate the ECE for p_θ by reversing the transformation, or we can notice that, if $ECE*$ is the transformed ECE, then $2ECE* \geq ECE$.

Finally, it is important to note that an estimate of an ECE is stochastic and a proper analysis must account for its uncertainty [20].

References

[1] Andreas Christmann and Ingo Steinwart. “Universal Kernels on Non-Standard Input Spaces”. In: *Advances in Neural Information Processing Systems*. Ed. by J Lafferty et al. Vol. 23. Curran Associates, Inc., 2010.

[2] A P Dawid. “The well-calibrated Bayesian”. en. In: *J. Am. Stat. Assoc.* 77.379 (Sept. 1982), pp. 605–610.

[3] Clara Fannjiang et al. “Conformal prediction for the design problem”. In: (Feb. 2022). arXiv: `2202.03613 [cs.LG]`.

[4] Shai Feldman, Stephen Bates, and Yaniv Romano. “Improving conditional coverage via orthogonal quantile regression”. In: *Adv. Neural Inf. Process. Syst.* 34 (2021), pp. 2060–2071.

[5] Adam Fisch, Tommi Jaakkola, and Regina Barzilay. “Calibrated Selective Classification”. In: (Aug. 2022). arXiv: `2208.12084 [cs.LG]`.

[6] Pierre Glaser et al. “Fast and Scalable Score-Based Kernel Calibration Tests”. In: ().

[7] Will Grathwohl et al. “Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One”. In: (Dec. 2019). arXiv: `1912.03263 [cs.LG]`.

[8] Arthur Gretton et al. “A kernel two-sample test”. In: *J. Mach. Learn. Res.* 13 (2012), pp. 723–773.

[9] Alex Kendall and Yarin Gal. “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?” In: (Mar. 2017). arXiv: `1703.04977 [cs.CV]`.

[10] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. “Trainable Calibration Measures for Neural Networks from Kernel Mean Embeddings”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2805–2814.

[11] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles”. In: *Advances in Neural Information Processing Systems*. Ed. by I Guyon et al. Vol. 30. Curran Associates, Inc., 2017.

[12] Wesley Maddox et al. “A Simple Baseline for Bayesian Uncertainty in Deep Learning”. In: (Feb. 2019). arXiv: `1902.02476 [cs.LG]`.

[13] Charles Marx et al. “Modular Conformal Calibration”. In: (June 2022). arXiv: `2206.11468 [cs.LG]`.

[14] Alexandre Perez-Lebel, Marine Le Morvan, and Gaël Varoquaux. “Beyond calibration: estimating the grouping loss of modern neural networks”. In: (Oct. 2022). arXiv: `2210.16315 [cs.LG]`.

[15] John C Platt. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in Large-Margin Classifiers*. Ed. by Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf and Dale Schuurmans. 1999.

[16] Samuel Stanton, Wesley Maddox, and Andrew Gordon Wilson. "Bayesian Optimization with Conformal Prediction Sets". In: (Oct. 2022). arXiv: `2210.12496 [cs.LG]`.

[17] Dustin Tran et al. "Plex: Towards Reliability using Pretrained Large Model Extensions". In: (July 2022). arXiv: `2207.07411 [cs.LG]`.

[18] Gia-Lac Tran et al. "Calibrating Deep Convolutional Gaussian Processes". In: (May 2018). arXiv: `1805.10522 [stat.ML]`.

[19] Uddeshya Upadhyay et al. "Posterior Annealing: Fast Calibrated Uncertainty for Regression". In: (Feb. 2023). arXiv: `2302.11012 [cs.LG]`.

[20] Juozas Vaicenavicius et al. "Evaluating model calibration in classification". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Ed. by Kamalika Chaudhuri and Masashi Sugiyama. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 3459–3467.

[21] David Widmann, Fredrik Lindsten, and Dave Zachariah. "Calibration tests beyond classification". In: (Oct. 2022). arXiv: `2210.13355 [stat.ML]`.

[22] Andrew Gordon Wilson and Pavel Izmailov. "Bayesian deep learning and a probabilistic perspective of generalization". In: *Adv. Neural Inf. Process. Syst.* 2020-Decem.3 (2020).

[23] Andrew Gordon Wilson et al. "Deep Kernel Learning". In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Ed. by Arthur Gretton and Christian C Robert. Vol. 51. Proceedings of Machine Learning Research. Cadiz, Spain: PMLR, 2016, pp. 370–378.