JMLR: Workshop and Conference Proceedings 19 (2011) 809–811 24th Annual Conference on Learning Theory

Does an Efficient Calibrated Forecasting Strategy Exist?

Jacob Abernethy JAKE@CS.BERKELEY.EDU *UC Berkeley, Div. of Computer Science*

Shie Mannor SHIE@EE.TECHNION.AC.IL *Technion, Department of Electrical Engineering*

Editor: Sham Kakade, Ulrike von Luxburg

Abstract We recall two previously-proposed notions of *asymptotic calibration* for a forecaster making a sequence of probability predictions. We note that the existence of efficient algorithms for calibrated forecasting holds only in the case of binary outcomes. We pose the question: do there exist such efficient algorithms for the general (non-binary) case?

Review of Calibrated Forecasting

Glenn Brier, writing in the journal *Monthly Weather Review*, observed a challenge in assessing sequential probability forecasts (Brier, 1950):

> *Verification of weather forecasts has been a controversial subject for more than a half century. There are a number of reasons why this problem has been so perplexing to meteorologists and others but one of the most important difficulties seems to be in reaching an agreement on the specification of a scale of goodness for weather forecasts.*

In response, he proposed an objective scoring function which, if implemented by the forecaster, would lead to "calibrated" predictions. Yet a question which presumably did not occur to Brier is whether the latter *is even computationally feasible*. This question will be topic of the present note.

Precisely, we would like to know whether there exists an efficient algorithm for forecasting sequences of outcomes that is *asymptotically calibrated*. As the forecaster observes the sequence $y_1, y_2, \ldots$, from some given set K, she outputs a sequence of (potentially randomized) probability predictions $\mathbf{p}_1, \mathbf{p}_2, \ldots \in \Delta(K)$, where $\mathbf{p}_t$ may depend only on the past $y_1, y_2, \ldots, y_{t-1}$. Roughly speaking, a forecaster is asymptotically calibrated if her probability predictions match the outcome frequencies. Let $\tau_T(\mathbf{p}) := \{t : \mathbf{p}_t \approx \mathbf{p}\}$, and let $\delta_y \in \Delta(K)$ be a point mass distribution on y, then our forecaster is calibrated if, for all $\mathbf{p} \in \Delta(K)$,

$$\frac{\sum_{t \in \tau_T(\mathbf{p})} \delta_{y_t}}{|\tau_T(\mathbf{p})|} \to \mathbf{p} \quad \text{as} \quad T \to \infty.$$

In the above definition we have been intentionally vague about the meaning of $\mathbf{p}_t \approx \mathbf{p}$, as well as our notion of convergence in the limit. To be more precise we shall distinguish between two notions of calibration: strong and weak.

Definition 1 *A function $g : \Delta(K) \to \mathbb{R}$ is called* lipschitz *if there exists some $c > 0$ for which $|g(\mathbf{p}) - g(\mathbf{q})| \leq c\|\mathbf{p} - \mathbf{q}\|_1$ for every $\mathbf{p}, \mathbf{q} \in \Delta(K)$. We shall call a forecaster* weakly

calibrated *if for every lipschitz "test function" g and any sequence $y_1, y_2, \ldots \in K$ we have*

$$\overbrace{\left\| \frac{1}{T} \sum_{t=1}^{T} g(\mathbf{p}_t)(\mathbf{p}_t - \delta_{y_t}) \right\|_1}^{CS_T(g) \text{ := calibration score w.r.t. } g} \to 0 \quad \text{almost surely as } T \to \infty. \tag{1}$$

Let $I_{\mathbf{p},\epsilon} : \Delta(K) \to \mathbb{R}$ be the indicator defined as $I_{\mathbf{p},\epsilon}(\mathbf{q}) := 1$ when $\|\mathbf{p} - \mathbf{q}\|_1 \leq \epsilon$ and $I_{\mathbf{p},\epsilon}(\mathbf{q}) := 0$ otherwise. A forecaster is called strongly calibrated *if equation* (1) *holds for the indicator test functions $g(\mathbf{q}) = I_{\mathbf{p},\epsilon}(\mathbf{q})$ for every $\mathbf{p} \in \Delta(K)$ and $\epsilon > 0$. Given a fixed ϵ, a forecaster is* ϵ-strongly calibrated *if, for every $\mathbf{p} \in \Delta(K)$, the calibration score $CS_T(I_{\mathbf{p},\epsilon}) \leq \epsilon$ for large enough T.*

It should be clear from the definitions that strong calibration implies weak, but it may not be as obvious that the reverse does not hold. It has been shown that for the binary case ($K = \{0, 1\}$) there exists a deterministic weakly calibrated forecaster (Kakade and Foster, 2008; Mannor et al., 2007), yet any strongly calibrated forecaster must be randomized. The latter is demonstrated by a well-known counter example of Dawid (1982). (For a complete survey on strongly calibrated forecasting, see Cesa-Bianchi and Lugosi (2006).)

The (strong or weak) calibration property leaves much to be desired as a measure of "performance" of a forecaster. For one, a forecaster can be calibrated yet completely inaccurate even on "easy" inputs: If the outcomes simply alternates as $(0, 1, 0, 1, 0, 1, \ldots)$ then the trivial forecaster predicting 0.5 achieves asymptotic calibration. Yet calibrated forecasting remains a useful tool in many circumstances, particularly given that it is robust to arbitrary and potentially adversarial inputs. It can be shown, for instance, that no-regret learning algorithms can be constructed from strongly-calibrated forecasters; the same trick can be applied towards a strategy for Blackwell's Approachability problem.

Open Problem: Can We Calibrate Efficiently?

The existence of a calibrated forecaster has been established for some time (strong (Foster and Vohra, 1998) and weak (Kakade and Foster, 2008)) in most cases via construction. Unfortunately, these constructions have typically been characterized by very greedy methods that give rise to inefficient algorithms. The most common approach is to take the probability space $\Delta(K)$ and cover it with an ϵ-grid and, for each grid point p, the algorithm maintains some statistic. To achieve ϵ-strongly-calibrated forecaster, the algorithm will potentially have to process the entire grid *for each prediction p_t*.

This story may have a happy ending. Recent results suggest that calibration may be achieved via efficient methods. Mannor et al. (2007) developed a weakly calibrated forecasting algorithm that requires constant time and space for each prediction. In the present year's COLT proceedings, Abernethy et al. (2011) developed an ϵ-strongly calibrated forecasting algorithm which requires $O(\log \frac{1}{\epsilon})$ time yet $O(\frac{1}{\epsilon})$ space per prediction.

What's the catch? Each of the above algorithm are applicable only to *binary* forecasting and do not extend to more general K. Mannor and Stoltz (2010) have recently introduced work addressing the case where $|K| > 2$, utilizing Blackwell approachability, to obtain ϵ-strong calibration with time and space complexity that behaves like $O(1/\epsilon^{|K|})$. As described

in the table below, very little progress has been made towards efficient algorithms that achieve calibration in general.

	Strong	Weak
Binary ($\|K\| = 2$)	$O(\log \frac{1}{\epsilon})$ time, $O(\frac{1}{\epsilon})$ space	$O(1)$ time/space
Finite-alphabet ($\|K\| = n > 2$)	$O(1/\epsilon^{\|K\|})$ time/space	?

We use the term "efficient" to denote an algorithm whose per-step complexity and memory requirements are poly in $|K|$ and poly-logarithmic in $(1/\epsilon)$. Our concrete questions are:

1. Is there an efficient time and memory algorithm for ϵ-strong calibration for $|K| = 2$?
2. Is there an efficient time and memory algorithm for weak calibration for any $|K|$?
3. Is there an efficient time and memory algorithm for strong calibrationfor any $|K|$?

Our best guess is that such an algorithm likely exists, for both the weak and strong case, in particular because we have not placed any restrictions on the rate at which the algorithm must achieve the calibration objective. On the other hand, the previously-discovered tricks which lead to efficient calibrated forecasters may be very special to the binary case and we would not be surprised if no such efficient algorithms exist when $|K| > 2$.

References

J. Abernethy, P.L. Bartlett, and E. Hazan. Blackwell approachability and low-regret learning are equivalent. In *Proceedings of the 24th Annual Conference on Learning Theory*, 2011.

G.W. Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78 (1):1–3, 1950. ISSN 1520-0493.

Nicolo Cesa-Bianchi and Gàbor Lugosi. *Prediction, Learning, and Games*. 2006. ISBN 0521841089, 9780521841085.

A. Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77:605–613, 1982.

D. P Foster and R. V Vohra. Asymptotic calibration. *Biometrika*, 85(2):379, 1998.

S.M. Kakade and D.P. Foster. Deterministic calibration and nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008.

S. Mannor and G. Stoltz. A geometric proof of calibration. *Mathematics of Operations Research*, 35(4):721–727, 2010. URL `http://webee.technion.ac.il/people/shie/public/papers/J_MShimkin03Bayes.pdf`.

S. Mannor, J.S. Shamma, and G. Arslan. Online calibrated forecasts: Memory efficiency versus universality for learning in games. *Machine Learning*, 67(1):77–115, 2007. ISSN 0885-6125.