Faster online calibration without randomization: interval forecasts and the power of two choices

Chirag Gupta and Aaditya Ramdas

Kiho Park

parkkiho@uchicago.edu

1 Introduction

A number of machine learning and statistics applications rely on probabilistic predictions. We study the problem of producing probabilistic forecasts for binary events, that are calibrated without any assumptions on the data-generating process.

1.1 Calibration games and ϵ-calibration

Calibration-Game-I (classical) models the problem as a game between a forecaster and nature. At time $t = 1, 2, \cdots$, Forecaster produces a randomized forecast $u_t \in \Delta([0,1])$. Here, $\Delta(S)$ denotes the set of probability distributions over the set S. Nature plays $v_t \in \Delta(\{0,1\})$. Forecaster predicts $p_t \sim u_t$. Nature reveals $y_t \sim v_t$. At time $T > 1$, the prior activities $(u_t, v_t, p_t, y_t)_{t=1}^{T-1}$ are known to both players. The goal of the forecaster is to appear calibrated. Nature wishes to prevent the forecaster from appearing calibrated. such a nature is typically referred to as an adaptive adversary.

Let $\epsilon > 0$ be a discretization or tolerance level. We assume that $\epsilon = 1/2m$ for some integer $m \geq 2$. Consider the ϵ-packing of $[0,1]$ given by the m intervals $I_1 = [0, 1/m), I_2 = [1/m, 2/m), \cdots, I_m = [1-1/m, 1]$. At time t, the forecaster makes a forecast corresponding to the mid-points of these intervals: $p_t \in \{M_1 := 1/2m, M_2 := 3/2m, \cdots, M_m := 1 - 1/2m\}$. Define $N_i^T := |\{t \leq T : p_t = M_i\}|$ and $p_i^T = \frac{1}{N_i^T} \sum_{t \leq T : p_t = M_i} y_t$ if $N_i^T > 0$, $p_i^T = M_i$ otherwise. Following Foster (1999), we define the (ℓ_1-)calibration error as the weighted sum of the prediction errors for each possible forecast:

$$\sum_{i=1}^{m} \frac{N_i^T}{T} |M_i - p_i^T| = \sum_{i=1}^{m} \left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{p_t = M_i\}(M_i - y_t) \right|$$

By the martingale convergence theorem, for a fixed i,

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{p_t = M_i\}(M_i - y_t) \xrightarrow{\text{a.s.}} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\mathbb{1}\{p_t = M_i\}(M_i - y_t)]$$

Thus, we say that the forecaster is ϵ-calibrated if for any strategy of nature, we have:

$$\sum_{i=1}^{m} \left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\mathbb{1}\{p_t = M_i\}(M_i - y_t)] \right| = \epsilon + f(T) = \epsilon + o(1)$$

We call the term $f(T)$ as calibration rate. In this paper, we are interested in the asymptotic dependence of $f(T)$ as $T \to \infty$, holding ϵ as a fixed problem parameter on which f may depend arbitrarily.

1.2 Related work and our contributions

A number of papers have proposed ϵ-calibrated forecasting algorithms which guarantee $f(T) = O(1/\sqrt{T})$. In Theorem 3, we show that the $O(1/\sqrt{T})$ rate by these algorithms is tight. Recently, it was shown that the

worst-case overall calibration rate ($\epsilon + f(T)$) is at least $\Omega(T^{-0.472})$. Observe that the overal calibrate rate must be $\Omega(T^{-1/2})$ if nature simply plays a non-adaptive Bernoulli strategy.

Foster (1999) showed that calibration is a Blackwell approachability instance, and while the rate $f(T) = \Omega(1/\sqrt{T})$ has not been formally established earlier, it is the rate one expects. Instead, the community has looked to establish positive results for alternative notions.

In our work, we retain the classical definition of ϵ-calibartion but change the calibration game. In Calibration-Game-II, also called the 'Power-Of-Two-Choices' (POTC) game, the forecaster reveals two forecasts $p_{t0}, p_{t1} \in [0,1]$, such that $p_{t0} \leq p_{t1}$ and $|p_{t1} - p_{t0}| \leq 2\epsilon$. There is no randomization, and nature knows the two forecasts. Nature reveals $y_t \in \{0,1\}$. If $y_t = 1$, set $p_t = p_{t1}$; else set $p_t = p_{t0}$.

2 Motivation for the POTC calibration game

Calibration-Game-II (POTC) is motivated by two rich fields of literature: imprecise probability and the power of two choices.

2.1 practical perspective via imprecise probability

In the area, imprecise probability, one is typically not restricted to work with single, unique probability measures, but instead the axioms of probability are relaxed, and added flexibility is provided in order to work with upper and lower probability measures.

In the context of our problem, instead of saying the the probability of rain is 0.3, a forecaster is allowed to say $0.3 \pm \epsilon$. From a practical perspective, this type of interval forecast arguably has almost the same utility and interpretability to a layman as the corresponding point forecast. The upper and lower forecasts is standard in game-theoretic probability.

Remarkably, this small and seemingly insignificant change in reporting leads to a huge change in our ability to achieve calibration. Of course, we must figure out how to judge the quality of such an interval forecast. We replace M_i with the projection of y_t onto $A_t := [p_{t0}, p_{t1}]$.

Imprecise probability has also made an intriguing appearance in the simpler setting of achieving calibration in offline binary classification in the presence of i.i.d. data. "Venn predictors" are a class of distribution-free algorithms that produce imprecise probability forecasts. One can view our work as extending the use of such imprecise interval forecasts to the online calibration setting with adversarial data.

2.2 The varied applications of the power of two choices

The power of two choices (POTC) refers to a remarkable result by Azar et al. (1994) for the problem of load balancing. Suppose n balls are placed independently and uniformly at random into n bins. It can be shown that with high probability, the maximum number of balls in a bin (the maximum 'load') will be $\widetilde{\Theta}(\log n)$. Consider a different setup where the balls are placed sequentially, and for each ball, two bin indices are drawn uniformly at random and offered to a 'load-balancer' who gets to decide which of the two bins to place the ball in. At each step, the load-balancer place the ball in the bin with lesser load. It turns out that with this strategy the maximum load drops exponentially to $\widetilde{\Theta}(\log \log n)$.

The POTC result has led a number of applications. In this paper, we find yet another intriguing phenomenon involving the POTC, this time in the context of calibration. We modify the classical setup of

calibration to the POTC setup, by offering the forecaster two nearby choices. We show that this chanbe accords the forecaster with significant power, enabling fasater calibration, even without randomization.

3 Main results: algorithm and analysis

Consider the POTC game. The forecaster can play either $(p_{t0}, p_{t1}) = (M_i, M_i)$ or $p_{t0}, p_{t1} = (M_i, M_{i+1})$ for some i. For $i \in [m]$ and $t \geq 1$, define $l_i = (i-1)/m$, $r_i = i/m$, $N_i^t = |\{\mathbb{1}\{p_s = M_i\} : s \leq t\}|$, $p_i^t = \sum_{s=1}^t y_s \mathbb{1}\{p_s = M_i\}/N_i^t$ if $N_i^t > 0$, $p_i^t = M_i$ otherwise, $d_i^t = l_i - p_i^t$ (deficit), and $e_i^t = p_i^t - r_i$ (excess).

3.1 Forecasting algorithm

The algorithm is a variant of the one proposed by Foster (1999).

Algorithm 1 POTC-Cal: Algorithm for forecaster in Calibration-Game-II

At time $t = 1$, play $(p_{10}, p_{11}) = (M_1, M_1)$. Thus $p_1 = M_1$.
At time $t+1$ $(t \geq 1)$
if condition A: there exists an interval $i \in [m]$ such that $d_i^t \leq 0$ and $e_i^t \leq 0$ **then**
  play (M_i, M_i) for any i that verifies condition A (i.e., $p_{t+1} = M_i$)
else
  condition B: there exists an interval $i \in [m-1]$ such that $e_i^t > 0$ and $d_{i+1}^t > 0$ must be satisfied.
  Play (M_i, M_{i+1}) for any index i that verifies condition B
  (i.e., $p_{t+1} = M_i$ if $y_{y+1} = 0$ and $p_{t+1} = M_{i+1}$ if $y_{y+1} = 1$)
end if

Lemma 1. *For any $t \geq 1$, if condition A is not satisfied, then condition B must be satisfied.*

3.2 Analysis of POTC-Cal

Lemma 2. *Suppose condition A was satisfied at time $t+1$ and the forecast was $p_{t+1} = M_i$. Then, $N_i^{t+1} \max(d_i^{t+1}, e_i^{t+1}) \leq 1$*

Lemma 3. *Suppose condition A was not satisfied at time $t+1$ and the forecast was $p_{t+1} = M_i$, following condition B. Then $N_i^{t+1} \max(d_i^{t+1}, e_i^{t+1}) \leq \max(N_i^t \max(d_i^t, e_i^t), 1)$.*

Theorem 1. *POTC-Cal satisfies, at any time $T \geq 1$, for any strategy of nature,*

$$\sum_{i=1}^{m} \left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{p_t = M_i\}(M_i - y_t) \right| \leq \epsilon + \frac{m}{T} = \epsilon + O(1/T)$$

3.3 Generalization to bounded outputs

If the output is bounded instead of binary, then POTC-Cal can be modified as follows. Other steps are same as the original POTC-Cal. However, p_i^T is the mean of the v_t values instead of y_t. Also, the forecaster plays $p_t = p_{t0}$ if $v_t \leq r_i$ and $p_t = p_{t1}$ if $v_t > r_i$. Lemmas 2 and 3 also hold, so the $O(1/T)$ rate in Theorem 1 also holds.

4 $\Omega(1/\sqrt{T})$ lower bound for the classical calibration game

Calibration-Game-I can be vied as a repeated game with vector-valued payoffs/rewards. Such games are now commonly referred to as Blackwell approachability games. We use it to prove the lower bound. We denote the action space of the forecaster as $\mathcal{X} = \{M_1, M_2, \cdots, M_m\}$ and that of nature as $\mathcal{Y} = \{0, 1\}$. The random plays of the forecaster line in $\Delta(\mathcal{X})$ which is a probability simplex in $\mathbb{R}^m$.

4.1 Calibration as an instance of Blackwell approachability

Suppose the actions of the forecaster and nature give a reward $r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^m$ where $a = r(p \in \mathcal{X}, y \in \mathcal{Y}) \in \mathbb{R}^m$ with $a_i = \mathbb{1}\{p = M_i\} \cdot (M_i - y)$. Let $\overline{a}^T := \sum_{t=1}^T a(p_t, y_t)/T$. Then, $\overline{a}_i^T = \sum_{t=1}^T \mathbb{1}p_t = M_i(M_i - y_t)/T$. The ϵ-calibration condition is equivalent to $\lim_{T\to\infty} \mathbb{E}[\overline{a}^T] \in B_\epsilon$ where B_ϵ is the ℓ_1-ball with radius ϵ. Can the forecaster play a strategy to ensure it? Blackwell (1956) termed this as a quesiton of 'approachability' of $\mathbb{E}[\overline{a}^T]$ towards B_ϵ.

Theorem 2 (Corollary to Theorem 3 by Blackwell (1956))**.** *Assume the same setup as calibration-Game-I, but the players receive a vector-valued reward r. The forecaster can ensure that $\lim_{T\to\infty} \mathbb{E}[\overline{a}^T] \in B_\epsilon$ if and only if for every $v \in \Delta(\mathcal{Y}) = [0, 1]$ of natrue, there exists some $u \in \Delta(\mathcal{X})$ such that the expected reward belongs to B_ϵ:*

$$\mathbb{E}_{p\sim u, y\sim v}[r(p, y)] = \sum_{i=1}^m u_i((1 - v) \cdot r(M_i, 0), v \cdot r(M_i, 1)) \in B_\epsilon$$

If the forecaster can respond to every play by nature and unsure that the expected reward lies in B_ϵ, then the forecaster can ensure that $\overline{a}^T$ approaches B_ϵ in the sequential game. Theorem 2 canb e read as response-satisfiability $\iff$ approachability.

Proposition 1. *The forecaster can exhibit response satisfiability. Thus the forecaster playing Calibration-Game-I can ensure $\lim_{T\to\infty} \mathbb{E}[\overline{a}^T] \in B_\epsilon$, and be ϵ-calibrated.*

4.2 A lower bound for the ϵ-calibration error rate

Theorem 3. *A forecaster playing Calibration-Game-I against an adversarial nature cannot achieve ϵ-calibration at a rate faster than $O(1/\sqrt{T})$. That is, for every strategy of the forecaster, there is a strategy of nature that ensures*

$$\sum_{i=1}^m \left| \frac{1}{T} \sum_{t=1}^T \mathbb{E}[\mathbb{1}\{p_t = M_i\}(M_i - y_t)] \right| = \epsilon + \Omega(1/\sqrt{T})$$

We use the following result which implies Theorem 3 once the desired conditions are satisfied by B_ϵ.

Theorem 4 (Theorem 6.ii by Mannor and Perchet (2013))**.** *Let $\mathcal{C}$ be a closed convex set that is (i) minimal approachable, and (ii) mixed approachable. Then $\mathcal{C}$ cannot be approached at a rate faster than $O(1/\sqrt{T})$.*

4.3 Minimal approachability

For a point $u \in \mathbb{R}^m$ and a convex set $K \subseteq \mathbb{R}^m$, define the distance of u from K as $d_K(u) = \inf_{u'\in K} \|u - u'\|_2$. For any $\lambda > 0$, a convex set $K' \subseteq K$ is said to be a λ-shrinkage of K if $\{u : d_{K'}(u) \leq \lambda\} \subseteq K$.

Definition 1. *A set K is minimal approachable if K is approachable, but no λ-shrinkage of K is approachable.*

Lemma 4. *The set B_ϵ is minimal approachable.*

4.4 Reducing approachability to scalar-valued games

Consider the approachability of B_ϵ with respect to individual directions, represented by arbitrary vectors $q \in \mathbb{R}^m$. Let $c \in B_\epsilon$ be such that q belongs to the normal cone of B_ϵ at c, that is $\langle c, q\rangle = \sup_{c' \in B_\epsilon} \langle c', q\rangle = \epsilon\|q\|_\infty$. We call such a pair (c, q) as 'admissible'. Consider the following one-shot min-max game defined for every admissible (c, q):

$$\begin{aligned}
\mathrm{Val}(c, q) &= \min_{u \in \Delta(\mathcal{X})} \max_{v \in [0,1]} \langle \mathbb{E}_{p\sim u, y\sim v}[r(p, y)] - c, q\rangle \\
&= \min_{u \in \Delta(\mathcal{X})} \max_{v \in [0,1]} \left(\sum_{i=1}^m u_i q_i((1-v)M_i + v(M_i - 1)) - \langle c, q\rangle \right) \\
&= \min_{u \in \Delta(\mathcal{X})} \max_{v \in [0,1]} \left(\sum_{i=1}^m u_i q_i(M_i - v) - \epsilon\|q\|_\infty \right)
\end{aligned}$$

The following argument illustrates the relationship between the $\mathrm{Val}(c, q)$ games and the B_ϵ-approachability game.

Suppose the forecaster can guarantee 'one-shot approachability', that is, there exists a fixed $u^* \in \Delta(\mathcal{X})$ such that for every $v \in [0, 1]$, $\mathbb{E}_{p\sim u^*, y\sim v}[r(p, y)] \in B_\epsilon$. If follows that for every admissible (c, q), $\mathrm{Val}(c, q) \le 0$. This observation does not hold in the reverse direction. However, Blackwell (1956) showed that the result hoes hold in the reverse direction for the repeated approachability game.

Theorem 5 (Theorem 1 by Blackwell (1956))**.** *A convex set K is approachable if and only if for every admissible (c, q), $\mathrm{Val}(c, q) \le 0$.*

This condition has also been termed as halfspace-satisfiability. For our problem, the min-max game does not depend on c, once we replace $\langle c, q\rangle$ with $\epsilon\|q\|_\infty$. We thus simplify notation and index our games only by $q \in \mathbb{R}^m$:

$$\mathrm{Val}(q) = \min_{u \in \Delta(\mathcal{X})} \max_{v \in [0,1]} \left(\sum_{i=1}^m u_i q_i(M_i - v) - \epsilon\|q\|_\infty \right)$$

We know that B_ϵ is approachable and hence by Blackwell's result, halfspace-satisfiability must hold. That is, for every $q \in \mathbb{R}^m$, $\mathrm{Val}(q) \le 0$.

4.5 Pure* game and mixed approachability

In order to achieve a small value in the $\mathrm{Val}(q)$ game, the forecaster may play a randomized strategy, that is, $u^* \neq e_i$. On the other hand, since nature goes second, she has no incentive to randomize. In the following 'pure' game, the forecaster is also not allowed to randomize over his actions.

$$\mathrm{Val}^{\mathrm{P}}(q) := \min_{p \in \mathcal{X}} \max_{y \in \{0,1\}} \langle r(p, y), q\rangle - \epsilon\|q\|_\infty$$

Let us refer to this as the pure game, and the game in Section 4.4 as the mixed game. Characterizing a situation where the fast rate $O(1/n)$ is unachievable requires considering another game, whose value lies between the pure and mixed games. Define

$$\mathcal{X}^* = \{p \in \mathcal{X} : p \in \text{support}(u^*), \text{ where } u^* \text{ is some optimal mixed strategy for the forecaster}\}$$

$$\mathcal{Y}^* = \{y \in \{0,1\} : y \in \text{support}(v^*), \text{ where } v^* \text{ is some optimal mixed strategy for nature}\}$$

Then define the pure* game and its value as follows,

$$\text{Val}^*(q) := \min_{p \in \mathcal{X}^*} \max_{y \in \mathcal{Y}^*} \langle r(p,y), q\rangle - \epsilon\|q\|_\infty$$

It can be shown that for any q, $\text{Val}(q) \le \text{Val}^*(q) \le \text{Val}^{\text{p}}(q)$.

Definition 2. *An approachable set is said to be mixed approachable if there exists a $q \in \mathbb{R}^m$ such that while $\text{Val}(q) = 0$, $\text{Val}^*(q) > 0$.*

Lemma 5. *There exists a $q \in \mathbb{R}^m$ such that $\text{Val}^*(q) > \text{Val}(q) = 0$. Thus B_ϵ is mixed approachable.*

4.6 Relationship to previous lower bounds for calibration

A strategy of nature was recently constructed to ensures that the calibration error of any forecaster playing Calibration-Game-I satisfies:

$$\sum_{i=1}^{m} \left| \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\mathbb{1}\{p_t = M_i\}(M_i - y_t)] \right| = \Omega(T^{-0.472})$$

Suppose the worst-case calibration error is $g(T) = \epsilon + f(T)$, where f may depend on ϵ. The result means $g(T) = \Omega(T^{-0.472})$, whereas Theorem 3 shows tha t$f(T) = \Omega(T^{-0.5})$. It is unclear whether the final bound will be better or worse than $\Omega(T^{-0.472})$.

A trivial proof for $g(T) = \Omega(T^{-0.5})$ can be obtained by considering a 'Bernoulli strategy' for nature: at each time step, nature plays $y_t \sim \text{Bernoulli(p)}$ for some fixed $p \in [0,1]$ unknown to the forecaster. However, this Bernoulli strategy does not lead to a proof for $f(T) = \Omega(T^{-0.5})$ because there exists a forecaster who is ϵ-calibrated and achieves a $O(\text{poly}(\log T)/T)$ rate on the Bernoulli strategy.

5 Summary

This paper connects three rich areas of the literature in a natural way: online calibration, the power of two choices, and imprecise probability. In summary, we show that by allowing the forecaster to output a deterministic short interval of probabilities, we can achieve a faster rate of $O(1/T)$ for ϵ-calibration against a fully adaptive adversarial reality who presents the binary outcome after observing the interval forecast.

Several open questions remain, since we open a rather new line of investigation. We mention two: (a) lower bounds for our setting are unknown, and (b) we don't know if models providing more than two choices could possibly improve the rate further. We suspect that $1/T$ is the optimal rate since it corresponds to constant cumulative calibration error. Finally, it would be interesting to (c) figure out multidimensional analogs of our paper.

References

Yossi Azar, Andrei Z Broder, Anna R Karlin, and Eli Upfal. Balanced allocations. In *Symposium on Theory of Computing*, 1994.

David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956.

Dean P Foster. A proof of calibration via Blackwell's approachability theorem. *Games and Economic Behavior*, 29(1-2):73–78, 1999.

Shie Mannor and Vianney Perchet. Approachability, fast and slow. In *Conference on Learning Theory*, 2013.

6 Question

1. page 9, proof of Theorem 1, a means action? I think it is not defined before. It means choosing the interval i? or choosing $p_{t+1} = M_i$? I think it means $p_{t+1} = M_i$ because of the proof.