Complexity-Based Approach to Calibration with Checking Rules

D. Foster, A. Rakhlin, K. Sridharan, A. Tewari
Fig. 1. Calibration curves for expert (Experiment 1) and amateur (Experiment 2) players (numbers in parentheses indicate number of observations).

A Must-Have Slide

**Philadelphia Weather**

<table>
<thead>
<tr>
<th></th>
<th>Tonight</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
<th>Fri</th>
<th>Sat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Jun 20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partly Cloudy</td>
<td>PM T-Storms</td>
<td>Isolated T-Storms</td>
<td>Scattered T-Storms</td>
<td>Isolated T-Storms</td>
<td>Isolated T-Storms</td>
<td></td>
</tr>
<tr>
<td>83°</td>
<td>86°</td>
<td>82°</td>
<td>86°</td>
<td>84°</td>
<td></td>
<td></td>
</tr>
<tr>
<td>67°</td>
<td>71°</td>
<td>71°</td>
<td>69°</td>
<td>66°</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Calibration for individual sequences

- World of stochastic modeling
- World of individual sequences
Calibration for individual sequences

Assume $\{P_{\theta}\}$
Data $X_1, \ldots, X_n \sim P_{\theta}$
Estimate $\hat{\theta}$
Act as if $\hat{\theta}$ is correct
Calibration for individual sequences

Assume \( \{P_\theta\} \)
Data \( X_1, \ldots, X_n \sim P_\theta \)
Estimate \( \hat{\theta} \)
Act as if \( \hat{\theta} \) is correct

No assumption on source
Data \( X_1, \ldots, X_n \)
Calibrate probabilities
Act as if these are true
Why calibration?

Calibration can be used as an intermediate step for prediction and decision-making.

In some sense, if you can calibrate, then you “know” the mixed strategy of Nature.

- Convergence to Nash and Correlated Equilibria
- Blackwell Approachability
- Regret minimization
Why introduce checking rules??
Why introduce checking rules??

Consider the sequence $0,1,0,1,0,1,...$
Why introduce checking rules??

Consider the sequence $0,1,0,1,0,1,...$

forecast of 0.5 for is calibrated! Oops...
Why introduce checking rules??

Consider the sequence 0,1,0,1,0,1,...

forecast of 0.5 for is calibrated! Oops...

Calibration only cares about long-run frequency...
it is the minimum requirement for the forecaster
Why introduce checking rules??

Consider the sequence $0, 1, 0, 1, 0, 1, ...$

forecast of 0.5 for is calibrated! Oops...

Calibration only cares about long-run frequency... it is the minimum requirement for the forecaster

... but we can ask for more!
Setup
For round $t = 1, \ldots, T$
For round $t = 1, \ldots, T$
the player chooses a mixed strategy $q_t \in \Delta(\Delta_k)$
Setup

For round $t = 1, \ldots, T$

the player chooses a mixed strategy $q_t \in \Delta(\Delta_k)$

the adversary picks outcome $x_t \in \{e_1, \ldots, e_k\}$
Setup

For round $t = 1, \ldots, T$

the player chooses a mixed strategy $q_t \in \Delta(\Delta_k)$

the adversary picks outcome $x_t \in \{e_1, \ldots, e_k\}$

the player draws $f_t \in \Delta_k$ from $q_t$ and observes outcome $x_t$
Setup

History: \( z_t = ((f_1, x_1), \ldots, (f_{t-1}, x_{t-1})) \)

The set of all possible histories \( \mathcal{Z} = \bigcup_{t=1}^{T} (\Delta_k \times E_k)^t \)

Definition

A forecast-based checking rule is a binary-valued function
\[ c : \mathcal{Z} \times \Delta_k \mapsto \{0, 1\} \]

Calibration metric:
\[ R_T = \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^{T} c(z_t, f_t) \cdot (f_t - x_t) \right\| \]

where \( \zeta \) is a class of checking rules
Examples

\[ R_T = \sup_{c \in \zeta} \left| \frac{1}{T} \sum_{t=1}^{T} c(z_t, f_t) \cdot (f_t - x_t) \right| \]
Examples

\[ R_T = \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^{T} c(z_t, f_t) \cdot (f_t - x_t) \right\| \]

Example 1

Classical \( \epsilon \)-calibration:

\[ \zeta = \left\{ c_p(z_t, f_t) = 1_{\{\|f_t - p\| \leq \epsilon\}} : p \in \Delta_k \right\} \]
Examples

\[ R_T = \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^{T} c(z_t, f_t) \cdot (f_t - x_t) \right\| \]

Example 1

Classical \( \epsilon \)-calibration:

\[ \zeta = \left\{ c_p(z_t, f_t) = 1_{\left\| f_t - p \right\| \leq \epsilon} : p \in \Delta_k \right\} \]

Example 2

Let \( \hat{\theta}_{\theta, t} \) = forecast made by a probabilistic model \( P_{\theta} \)

\[ \zeta = \left\{ c_{\theta, p}(z_t, f_t) = 1_{\left\| \hat{\theta}_{\theta, t} - p \right\| \leq \epsilon} : p \in \Delta, \theta \right\} \]

will test if the model \( P_{\theta} \) is a much better fit to the data than \( f_t \)
Value of the Game

\[
\mathcal{V}_T^\theta(\zeta) = \inf_{q_1} \sup_{x_1} \mathbb{E}_{f_1 \sim q_1} \ldots \inf_{q_T} \sup_{x_T} \mathbb{E}_{f_T \sim q_T} \left[ 1 \left\{ \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^{T} c(z_t, f_t) \cdot (f_t - x_t) \right\| > \theta \right\} \right]
\]
Main Results
Main Results

Lemma

\[ \nu^\theta_T(\zeta) \leq 4 \sup_{x,p^\delta} \mathbb{P}_\varepsilon \left( \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^{T} \epsilon_t \ c(z_t^\delta(\varepsilon), p_t^\delta(\varepsilon)) \ x_t(\varepsilon) \right\| > \frac{\theta}{8} \right) \]
Main Results

Lemma

\[ \nu^\theta_T(\zeta) \leq 4 \sup_{x, p^\delta} \mathbb{P}_\epsilon \left( \sup_{c \in \zeta} \frac{1}{T} \sum_{t=1}^{T} \epsilon_t \ c(z_t^\delta(\epsilon), p_t^\delta(\epsilon)) \ x_t(\epsilon) \right) > \frac{\theta}{8} \]
Main Results

Lemma

\[ \nu_{T}^{\theta}(\zeta) \leq 4 \sup_{x, p^\delta} \mathbb{P}_\epsilon \left( \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^{T} \epsilon_t \ c(\mathbf{z}_t^\delta(\epsilon), p_t^\delta(\epsilon)) \ \mathbf{x}_t(\epsilon) \right\| > \frac{\theta}{8} \right) \]

Theorem

\[ \nu_{T}^{\theta}(\zeta) \leq 8 \ \mathcal{N}_{ch}(\zeta, T) \exp \left( -\frac{T\theta^2}{64 \ c_k} \right) \]
Main Results

Lemma

$$\mathcal{N}^\theta_T(\zeta) \leq 4 \sup_{x, p^\delta} \mathbb{P}_\epsilon \left( \sup_{c \in \zeta} \left\| \frac{1}{T} \sum_{t=1}^T \epsilon_t \, c(z_t^\delta(\epsilon), p_t^\delta(\epsilon)) \, x_t(\epsilon) \right\| > \frac{\theta}{8} \right)$$

Theorem

$$\mathcal{N}^\theta_T(\zeta) \leq 8 \, \mathcal{N}_{ch}(\zeta, T) \exp \left( -\frac{T \theta^2}{64 \, c_k} \right)$$
Families of checking rules
Families of checking rules

Finite classes of checking rules: $T^{-1/2}$ rate
Families of checking rules

Finite classes of checking rules: $T^{-1/2}$ rate

History invariant checking rules with VC dimension $VC(\zeta)$

$$c_k \sqrt{\frac{kVC(\zeta) \cdot \log(8/\eta) \log T}{T}}$$

and similar rate for bounded Littlestone’s dimension $Ldim(\zeta)$
Families of checking rules

Finite classes of checking rules: \( T^{-1/2} \) rate

History invariant checking rules with VC dimension \( VC(\zeta) \)

\[
C_k \sqrt{\frac{k \cdot VC(\zeta) \cdot \log(8/\eta) \log T}{T}}
\]

and similar rate for bounded Littlestone's dimension \( Ldim(\zeta) \)

Corollary: classical calibration with \( k \) actions and L-1 norm

\[
k^2 \sqrt{\frac{\log(T) \log(1/\eta)}{T}}
\]
Families of checking rules

Finite classes of checking rules: \( T^{-1/2} \) rate

History invariant checking rules with VC dimension \( VC(\zeta) \)

\[
c_k \sqrt{\frac{k \cdot VC(\zeta) \cdot \log(8/\eta) \log T}{T}}
\]

and similar rate for bounded Littlestone’s dimension \( Ldim(\zeta) \)

Corollary: classical calibration with \( k \) actions and L-1 norm

\[
k^2 \sqrt{\frac{\log(T) \log(1/\eta)}{T}}
\]

Arbitrary history-invariant checking rules: \( T^{-1/(k+1)} \) rate
More families of checking rules

- Time-dependent but history-invariant
- History represented by smaller set (e.g. bounded memory) (covering argument)
- Limited memory look-back
- Checking rules with bounded computation
For two actions, the rate for the *classical calibration* game is lower bounded

\[ \mathcal{V}_T^\theta \geq \mathbb{P} \left( \frac{1}{T} \sum_{t=1}^{T} \epsilon_t \geq 2\theta \right) = \Omega \left( \frac{1}{\sqrt{T}} \right) \]

This matches upper bound for classical calibration
Being calibrated can be viewed as a negative result!
A few remarks

Several very closely related ideas: calibration, no-internal regret, Blackwell’s approachability, online optimization.
Future work:

- Efficient Algorithms?
- Real-valued calibration rules (straightforward)
- Relation to Nash and Correlated Equilibria
- Relation to stochastic modeling