Blackwell Approachability and Calibrated Forecasting

Yen-Huan Li
yen-huan.li@epfl.ch

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

lions@epfl

MARIE CURIE ACTIONS

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

erc

EPFL

Recommended reading materials

1. N. Cesa-Bianchi and G. Lugosi. 2006. *Prediction, Learning, and Games.* Chapter 7. Cambridge Univ. Press.

2. S. Sorin. 2002. *A First Course on Zero-Sum Repeated Games.* Appendix B. Springer.

3. S. Mannor and G. Stoltz. 2010. A geometric proof of calibration. *Math. Oper. Res. 35*, 4, 721–727.

4. D. P. Foster and R. V. Vohra. 2011. Calibration: Respice, adspice, prospice.

Table of Contents

Problem of Binary Probability Forecasting

Can One Predict the Future?

Given an arbitrary sequence of binary digits

$$0\ 1\ 1\ 0\ 1\ 0\ 0\ 0\ 1\ 0\ 1,$$

can one predict the next digit, *without any probabilistic assumption*?

Weather Forecasting [3]

Think of '1' as "a rainy day", and '0' otherwise.

Given the weather history (say, of the last three years), *without any more information*, can one tell whether tomorrow will be rainy or not?

Probability Forecasting

What we typically see in a weather forecast is a probability forecast, e.g.,

THE CHANCE OF RAIN IS 80% TOMORROW.

Binary Probability Forecasting Protocol

For $t = 1, 2, 3, \ldots,$

- FORECASTER announces $p_t \in [0, 1]$.
- REALITY announces $y_t \in \{0, 1\}$, without knowing p_t.

Definition (Strategy)

A *strategy* of FORECASTER is a possibly random mapping:

$$p_t = f(p_1, y_1, p_2, y_2, \ldots, p_{t-1}, y_{t-1}, \xi_t),$$

where ξ_t is a random variable (r.v.). Similarly, a strategy of REALITY is a mapping

$$y_t = g(p_1, y_2, p_2, y_2, \ldots, p_{t-1}, y_{t-1}, \eta_t),$$

where η_t is another r.v. independent of ξ_t.

Calibrated Forecasting: Intuition

Intuitive Definition

A strategy of FORECASTER is *calibrated*, if for any strategy of REALITY and any $\rho \in [0, 1]$,

$$\frac{\sum_{t=1}^{T} \mathbb{I}_{\{y_t=1, p_t \approx \rho\}}}{\sum_{t=1}^{T} \mathbb{I}_{\{p_t \approx \rho\}}} \approx \rho.$$

Example

Suppose the weather forecast is calibrated. Then if the chance of rain is said to be 80%, a reasonable person should take an umbrella.

Obviously, this "definition" is not precise enough to work with. Some things to clarify:

1. What is the value of T?
2. What is the precise definition of "$p_t \approx \rho$"?
3. What is the precise definition of "the ratio $\approx \rho$"?

Calibrated Forecasting: Formalism (1/2)

Consider an ε-covering $\mathcal{Q}_\varepsilon := \{q_1, q_2, \ldots, q_{M_\varepsilon}\}$ of the interval $[0,1]$. Suppose p_t only takes value in $\mathcal{Q}_\varepsilon$. Define the empirical frequencies

$$f_m(T) := \frac{\sum_{t=1}^T \mathbb{I}_{\{p_t = q_m, y_t = 1\}}}{\sum_{t=1}^T \mathbb{I}_{\{p_t = q_m\}}}, \quad m = 1, 2, \ldots, M_\varepsilon,$$

whenever the denominator is non-zero, and $f_m(T) := q_m$ otherwise. We want

$$\limsup_{T \to \infty} |f_m(T) - q_m| \leq \varepsilon \quad \text{a.s..}$$

We would like to ignore those m's such that $\sum_{t=1}^T \mathbb{I}_{\{p_t = q_m\}} = o(T)$. A compact formulation is to require

$$\limsup_{T \to \infty} \sum_{m=1}^{M_\varepsilon} |f_m(T) - q_m| \left(\frac{\sum_{t=1}^T \mathbb{I}_{\{p_t = q_m\}}}{T} \right) \leq \varepsilon \quad \text{a.s..}$$

Some calculation leads to the equivalent definition of ε-calibration in the next slide.

Calibrated Forecasting: Formalism (2/2)

Definition (ε-calibration [3, 9])

Consider an ε-covering $\mathcal{Q}_\varepsilon := \{q_1, q_2, \ldots, q_{M_\varepsilon}\}$ of the interval $[0,1]$. A strategy of FORECASTER taking values in $\mathcal{Q}_\varepsilon$ is ε-*calibrated*, if for any strategy of REALITY,

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{m=1}^{M_\varepsilon} \sum_{t=1}^{T} \mathbb{I}_{\{p_t = q_m\}} \left| q_m - y_t \right| \leq \varepsilon, \quad \text{a.s.}$$

Definition (Calibration [3, 13])

A strategy of FORECASTER is *calibrated*, if for any strategy of REALITY, $\varepsilon > 0$, and $\rho \in [0,1]$,

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}_{\{|p_t - \rho| \leq \varepsilon\}} \left| p_t - \mathbf{y}_t \right| = 0, \quad \text{a.s..}$$

Other Applications and Implications

Computing a correlated equilibrium [8]

For almost every two-player repeated game, the set of correlated equilibria equals the set of limit empirical distribution of actions by calibrated learning.

Testing [17]

If the truth passes a *finite test* with probability ρ, then there exists an ignorant algorithm that passes the test with the same probability.

Contextual online learning [19]

If the spaces of side information and outcomes are locally compact, and the loss function is continuous and compact type, there exists a *universally consistent* prediction algorithm.

Approximating a Nash equilibrium [7]

There exists an uncoupled strategy, such that in every repeated game, the empirical distribution of all players' actions converges to an ε-Nash equilibrium almost surely.

Some Impossibility Facts

Binary Probability Forecasting Protocol

For $t = 1, 2, 3, \ldots,$

- FORECASTER announces $p_t = f(p_1, y_1, \ldots, p_{t-1}, y_{t-1}, \xi_t) \in [0, 1]$.
- REALITY announces $y_t = g(p_1, y_1, \ldots, p_{t-1}, y_{t-1}, \eta_t) \in \{0, 1\}$, without knowing p_t.

Fact 1: Addressing REALITY that knows p_t Is Impossible

If p_t is known, REALITY can choose $y_t \sim \mathrm{Bernoulli}(\varphi(p_t))$ independently for all t, using some function $\varphi : [0, 1] \to [0, 1]$ very different from the identity mapping.

Fact 2: Deterministic Forecasting Is Impossible [4, 9, 16]

If FORECASTER adopts a deterministic strategy, i.e., ξ_t is absent, then REALITY can get the exact value of p_t by computing $f(p_1, y_1, \ldots, p_{t-1}, y_{t-1})$.

⋆A Weaker Notion of Calibration

Definition (Weak calibration)

A strategy of FORECASTER is *weakly calibrated* with respect to a set $\mathcal{F}$ of functions $f : [0,1] \to \mathbb{R}$, if for every $f \in \mathcal{F}$,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f(p_t)\,|y_t - p_t| = 0, \quad \text{a.s.}$$

Facts

- There exists deterministic weakly calibrated strategies for FORECASTER, even against REALITY that knows p_t [12, 20].
- The weakly calibrated strategy in [12] followed by random rounding leads to an ε-calibrated strategy.
- Weak calibration is computationally hard in general [10].

Goal of This Lecture

Theorem ([6, 9, 13])

There exists an ε-calibrated forecasting strategy for FORECASTER.

Proof.

Cast the problem of calibrated forecasting as an approachability problem. Prove by Blackwell's approachability theorem. □

Theorem ([13])

There exists a calibrated forecasting strategy for FORECASTER.

Proof.

Apply the Borel-Cantelli lemma. □

Blackwell Approachability

Repeated Games with Vector Payoffs

Protocol of a Repeated Game with Vector Payoffs

For $t = 1, 2, \ldots,$

1. ALICE announces $a_t \in \mathcal{A}$.
2. BOB announces $b_t \in \mathcal{B}$, without knowing a_t.
3. ALICE observes the payoff vector $f(a_t, b_t) \in \mathbb{R}^p$.

Notes

1. We assume both $\mathcal{A}$ and $\mathcal{B}$ are of finite cardinalities.
2. A *strategy* of ALICE is a mapping $h_{\mathsf{A}} : (\mathcal{A} \times \mathcal{B})^* \to \Delta(\mathcal{A})$; Alice chooses a_t randomly according to $p_t = h_{\mathsf{A}}(a_1, b_1, \ldots, a_{t-1}, b_{t-1}) \in \Delta(\mathcal{A})$.
3. A strategy $h_{\mathsf{B}} : (\mathcal{A} \times \mathcal{B})^* \to \Delta(\mathcal{B})$ of BOB is defined similarly.
4. For any $p_t \in \Delta(\mathcal{A})$, $q_t \in \Delta(\mathcal{B})$, define $f(p_t, q_t) := \sum_{a \in \mathcal{A}, b \in \mathcal{B}} f(a, b) p_t(a) q_t(b)$.

(Blackwell) Approachability

Protocol of a Repeated Game with Vector Payoffs

For $t = 1, 2, \ldots,$

1. ALICE announces $a_t \in \mathcal{A}$.
2. BOB announces $b_t \in \mathcal{B}$, without knowing a_t.
3. ALICE observes the payoff vector $f(a_t, b_t) \in \mathbb{R}^p$.

Definition (Blackwell Approachability [1])

A set $\mathcal{X} \subseteq \mathbb{R}^p$ is *approachable*, if ALICE has a strategy such that, for any strategy of BOB,

$$\lim_{T \to \infty} \operatorname{dist}\left(\bar{f}_T, \mathcal{X}\right) = 0, \quad \text{a.s.},$$

where $\bar{f}_T := T^{-1} \sum_{t=1}^{T} f(a_t, b_t)$, and $\operatorname{dist}(v, \mathcal{X}) := \|v - \operatorname{proj}_{\mathcal{X}}(v)\|_2$.

Example: Learning with Expert Advice (1/2)

Protocol of Learning with Expert Advice

For $t = 1, 2, \ldots,$

1. EXPERT-i announces $\gamma_t(i) \in \Gamma$, $i = 1, 2, \ldots, n$.
2. LEARNER announces $\gamma_t \in \Gamma$.
3. REALITY announces $\omega_t \in \Omega$, *without knowing* γ_t.
4. LEARNER observes the losses $\lambda(\gamma_t, \omega_t)$ and $\lambda_t(\gamma_t(i), \omega_t)$ for all i.

Definition (No Regret)

We say that a (possibly random) strategy of LEARNER feels no regret, if

$$\lim_{T\to\infty} \frac{1}{T} \left(\sum_{t=1}^{T} \lambda(\gamma_t, \omega_t) - \sum_{t=1}^{T} \lambda(\gamma_t(i), \omega_t) \right) \leq 0, \quad \text{for all } i \text{ a.s.}$$

Example: Learning with Expert Advice (2/2)

Protocol of Learning with Expert Advice

For $t = 1, 2, \ldots,$

1. EXPERT-i announces $\gamma_t(i) \in \Gamma$, $i = 1, 2, \ldots, n$.
2. LEARNER announces $\gamma_t \in \Gamma$.
3. REALITY announces $\omega_t \in \Omega$, *without knowing* γ_t.
4. LEARNER observes the vector payoff $f_t := (\lambda(\gamma_t, \omega_t) - \lambda(\gamma_t(i), \omega_t))_{1 \le i \le n} \in \mathbb{R}^n$.

Definition (No Regret)

Let $\mathcal{X}$ be the negative orthant, i.e., $\mathcal{X} := (-\infty, 0]^n$.

We say that a (possibly random) strategy of LEARNER feels no regret, if $\mathcal{X}$ is approachable by LEARNER, i.e.,

$$\lim_{T \to \infty} \operatorname{dist}\left(\frac{1}{T}\sum_{t=1}^{T} f_t, \mathcal{X}\right) = 0, \quad \text{a.s.}$$

Blackwell's Approachability Theorem

Theorem (Blackwell's Approachability Theorem [1])

A closed convex set $\mathcal{X}$ is approachable, if and only if for any $q_B \in \Delta(\mathcal{B})$, there exists some $p_A \in \Delta(\mathcal{A})$, such that $f(p_A, q_B) \in \mathcal{X}$.

The only if part is obvious.

Proof of the Only If Part.

Suppose there exists some $q_{\mathsf{B}}^{\star} \in \Delta(\mathcal{B})$, such that $f(p_{\mathsf{A}}, q_{\mathsf{B}}^{\star}) \notin \mathcal{X}$ for all $q_{\mathsf{A}} \in \Delta(\mathcal{A})$. Then Bob can simply choose $q_t = q_{\mathsf{B}}^{\star}$ for all t. □

Tools (1/2): von Neumann's Minimax Theorem

Definition (Convex-Concave Function)

A function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is convex-concave, if $g(\cdot, y)$ is convex for any $y \in \mathcal{Y}$, and $g(x, \cdot)$ is concave for any $x \in \mathcal{X}$.

Theorem (von Neumann's Minimax Theorem [18])

Let $\mathcal{X} \subseteq \mathbb{R}^{d_1}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_2}$ be convex compact sets. Let $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be convex-concave. Then it holds that

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x, y) = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} g(x, y).$$

Proof of the If Part (1/5)

Lemma (Blackwell's Condition)

If for any $q_B \in \Delta(\mathcal{B})$, there exists some $p_A \in \Delta(\mathcal{A})$, such that $f(p_A, q_B) \in \mathcal{X}$, then for any $v \in \mathbb{R}^p$, there exists some $p_A \in \Delta(\mathcal{A})$, such that

$$\langle v - \mathrm{proj}_{\mathcal{X}}(v), f(p_A, p_B) - \mathrm{proj}_{\mathcal{X}}(v)\rangle \leq 0, \quad \textit{for all } p_B \in \Delta(\mathcal{B}).$$

Proof.

By the optimality condition, one has

$$\langle v - \mathrm{proj}_{\mathcal{X}}(v), x - \mathrm{proj}_{\mathcal{X}}(v)\rangle \leq 0, \quad \text{for all } x \in \mathcal{X}.$$

Apply von Neumann's minimax theorem, this implies

$$\begin{aligned}
&\min_{p_\mathsf{A}} \max_{p_\mathsf{B}} \langle v - \mathrm{proj}_{\mathcal{X}}(v), f(p_\mathsf{A}, p_\mathsf{B}) - \mathrm{proj}_{\mathcal{X}}(v)\rangle \\
&\quad = \max_{p_\mathsf{B}} \min_{p_\mathsf{A}} \langle v - \mathrm{proj}_{\mathcal{X}}(v), f(p_\mathsf{A}, p_\mathsf{B}) - \mathrm{proj}_{\mathcal{X}}(v)\rangle \\
&\quad \leq 0.
\end{aligned}$$

□

Proof of the If Part (2/5)

Define the average payoff:

$$\bar{f}_T := T^{-1} \sum_{t=1}^{T} f(a_t, b_t).$$

A Strategy of ALICE [1]

For every $t \in \mathbb{N}$, find some $p_t \in \Delta(\mathcal{A})$ such that

$$\left\langle \bar{f}_{t-1} - \mathrm{proj}_{\mathcal{X}}(\bar{f}_{t-1}), f(p_t, q) - \mathrm{proj}_{\mathcal{X}}(\bar{f}_{t-1}) \right\rangle \leq 0, \quad \text{for all } q \in \Delta(\mathcal{B}).$$

Announce $a_t \in \mathcal{A}$ randomly according to p_t.

Note

- Existence of p_t is guaranteed by Blackwell's condition.
- p_t can be found by solving the bilinear saddle point problem [11, 15]

$$p_t \in \arg\min_{p \in \Delta(\mathcal{A})} \max_{q \in \Delta(\mathcal{B})} \left\langle \bar{f}_{t-1} - \mathrm{proj}_{\mathcal{X}}(\bar{f}_{t-1}), f(p, q) - \mathrm{proj}_{\mathcal{X}}(\bar{f}_{t-1}) \right\rangle.$$

Proof of the If Part (3/5)

Observation

It suffices to consider the approachability of

$$\tilde{\mathcal{X}} := \mathcal{X} \cap \operatorname{conv}\left(\{f(a,b) | a \in \mathcal{A}, b \in \mathcal{B}\}\right).$$

Lemma

Define

$$M := \max_{a \in \mathcal{A}, b \in \mathcal{B}} \|f(a,b)\|_2 , \quad R := \max_{v \in \tilde{\mathcal{X}}} \|v\|_2 .$$

The strategy in the previous slide ensures:

1. $\mathsf{E}\left[\operatorname{dist}(\bar{f}_t, \mathcal{X})\right] \leq t^{-1/2}(M+R)$ *for all* t.
2. $\operatorname{dist}(\bar{f}_t, \mathcal{X})$ *converges to zero almost surely, as* $t \to \infty$.

Proof of the If Part (4/5)

Define $f_t := f(a_t, b_t)$, $\pi_t := \mathrm{proj}_{\mathcal{X}}(\bar{f}_t)$, and $\delta_t := \mathrm{dist}(\bar{f}_t, \mathcal{X})$.

We write

$$\begin{aligned}\delta_{t+1}^2 &\leq \left\| \bar{f}_{t+1} - \pi_t \right\|_2^2 \\ &= \left(\frac{t}{t+1}\right)^2 \delta_t^2 + \left(\frac{1}{t+1}\right)^2 \|f_{t+1} - \pi_t\|_2^2 + \frac{2t}{(t+1)^2} \left\langle \bar{f}_t - \pi_t, f_{t+1} - \pi_t \right\rangle .\end{aligned}$$

Denote by E_t the conditional expectation w.r.t. $(a_1, b_1, \ldots, a_t, b_t)$. By the Blackwell condition, we write

$$\begin{aligned}\mathsf{E}_t \left[\delta_{t+1}^2\right] &\leq \left(\frac{t}{t+1}\right)^2 \delta_t^2 + \left(\frac{1}{t+1}\right)^2 \|f_{t+1} - \pi_t\|_2^2 \\ &\leq \left(\frac{t}{t+1}\right)^2 \delta_t^2 + \left(\frac{1}{t+1}\right)^2 (M+R)^2.\end{aligned}$$

By induction, we obtain

$$\mathsf{E}\left[\delta_t^2\right] \leq \frac{(M+R)^2}{t}.$$

Tools (2/2): Doob's Martingale Convergence Theorem

Definition (Supermartingale)

A sequence of random variables $X_1, X_2, \ldots$ forms a supermartingale, if $\mathsf{E}\left[|X_t|\right] < +\infty$ for all t, and

$$\mathsf{E}\left[X_{t+1} | X_1, X_2, \ldots, X_t\right] \leq X_t.$$

Theorem (Doob's Maringale Convergence Theorem [5])

A non-negative supermartingale converges to an integrable random variable almost surely.

Proof of the If Part (5/5)

Define the random variables

$$\xi_t := \delta_t^2 + (M+R)^2 \sum_{\tau > t} \frac{1}{(\tau+1)^2}.$$

Then

$$\mathsf{E}_t\left[\xi_{t+1}\right] \leq \left(\frac{t}{t+1}\right)^2 \delta_t^2 + \sum_{\tau > t} \frac{1}{(\tau+1)^2} \leq \xi_t,$$

implying that $(\xi_t)_{t\in\mathbb{N}}$ is a non-negative supermartingale.

Hence $(\delta_t)_{t\in\mathbb{N}}$ converges to an integrable random variable δ_∞ almost surely. Since $\mathsf{E}\left[\delta_t^2\right] \to 0$, we have $\delta_\infty = 0$ almost surely.

Blackwell's Approachability Theorem and an Achieving Strategy

Theorem (Blackwell's Approachability Theorem [1])

A closed convex set $\mathcal{X}$ is approachable, if and only if for any $q_B \in \Delta(\mathcal{B})$, there exists some $p_A \in \Delta(\mathcal{A})$, such that $f(p_A, q_B) \in \mathcal{X}$.

A Strategy of Alice [1]

For every $t \in \mathbb{N}$, find some $p_t \in \Delta(\mathcal{A})$ such that

$$\left\langle \bar{f}_{t-1} - \mathrm{proj}_{\mathcal{X}}(\bar{f}_{t-1}), f(p_t, q) - \mathrm{proj}_{\mathcal{X}}(\bar{f}_{t-1}) \right\rangle \leq 0, \quad \text{for all } q \in \Delta(\mathcal{B}).$$

Announce $a_t \in \mathcal{A}$ randomly according to p_t.

Remark

- Existence of p_t is guaranteed by Blackwell's condition.
- Computing p_t can be done by solving a bilinear saddle point problem.

Non-Asymptotic Result

We have proved the almost sure convergence of $\mathrm{dist}(\bar{f}_t, \mathcal{X})$ to zero. How fast can the convergence rate be?

Theorem (Essentially Exercise 7.23 in [2])

Assume that $\mathcal{X}$ satisfies the Blackwell condition. Define $\tilde{\mathcal{X}}$ as the intersection of $\mathcal{X}$ and the convex hull of $\{f(a,b) | a \in \mathcal{A}, b \in \mathcal{B}\}$.

There exists a strategy for ALICE, *such that for any strategy of* BOB *and any $T \in \mathbb{N}$, it holds with probability at least $1-\delta$ that*

$$\mathrm{dist}\left(\bar{f}_T, \mathcal{X}\right) \leq \frac{M+R}{\sqrt{T}} + 4\sqrt{\frac{M^2}{cT}\left[\log\left(\frac{1}{\delta}\right) + 2.5\right]} = O(T^{-1/2}),$$

where $M := \max_{a,b} \|f(a,b)\|_2$, $R := \max_{v \in \tilde{\mathcal{X}}} \|v\|_2$, and c is a positive constant.

Proof (1/3)

Define the average expected payoff:

$$\tilde{f}_T := T^{-1} \sum_{t=1}^{T} f(p_t, b_t).$$

An Improved Strategy of ALICE [2]

For every $t \in \mathbb{N}$, find some $p_t \in \Delta(\mathcal{A})$ such that

$$\left\langle \tilde{f}_{t-1} - \mathrm{proj}_{\mathcal{X}}(\tilde{f}_{t-1}), f(p_t, q) - \mathrm{proj}_{\mathcal{X}}(\tilde{f}_{t-1}) \right\rangle \leq 0, \quad \text{for all } q \in \Delta(\mathcal{B}).$$

Announce $a_t \in \mathcal{A}$ randomly according to p_t.

Note

- This is slightly different from the original strategy Blackwell proposed.
- Existence of p_t is guaranteed by Blackwell's condition.
- Computing p_t can be done by solving a bilinear saddle point problem.

Proof (2/3)

Define $\tilde{\pi}_t := \mathrm{proj}_{\mathcal{X}}(\tilde{f}_t)$ and $\tilde{\delta}_t := \mathrm{dist}(\tilde{f}_t, \mathcal{X})$.

Similarly as in the previous proof, we write

$$\begin{aligned}
\tilde{\delta}_{t+1}^2 &\leq \left\| \tilde{f}_{t+1} - \tilde{\pi}_t \right\|_2^2 \\
&= \left(\frac{t}{t+1}\right)^2 \tilde{\delta}_t^2 + \left(\frac{1}{t+1}\right)^2 \|f(p_{t+1}, b_t) - \tilde{\pi}_t\|_2^2 + \\
&\quad \frac{2t}{(t+1)^2} \left\langle \tilde{f}_t - \tilde{\pi}_t, f(p_t, b_t) - \tilde{\pi}_t \right\rangle \\
&\leq \left(\frac{t}{t+1}\right)^2 \tilde{\delta}_t^2 + \left(\frac{1}{t+1}\right)^2 (M+R)^2.
\end{aligned}$$

By induction, we get

$$\tilde{\delta}_t \leq \frac{M+R}{\sqrt{t}}.$$

Tool: Vector-Valued Martingale

Definition (Vector-Valued Martingale)

Consider a sequence of random vectors $(X_t)_{t\in\mathbb{N}}$. We say that $(X_t)_{t\in\mathbb{N}}$ is a *(vector-valued) martingale*, if for all t,

$$\begin{aligned}&\mathsf{E}[\|X_t\|_2] < +\infty,\\ &\mathsf{E}[X_{t+1}|X_1, X_2, \ldots, X_t] = X_t.\end{aligned}$$

Theorem (Azuma Inequality for Vector-Valued Martingales [14])

There exists a positive constant c, such that for any vector-valued martingale $(X_t)_{t\in\mathbb{N}}$ satisfying $\|X_{t+1} - X_t\|_2 \leq a_t$ for all t, it holds that

$$\mathsf{P}\left(\|X_{t+1} - X_1\|_2 \geq u\right) \leq \mathrm{e}^{2.5} \exp\left(-\frac{cu^2}{\sum_{\tau=1}^{t} a_\tau^2}\right),$$

for any $t \in \mathbb{N}$ and $u > 0$.

Proof (3/3)

Recall that $\bar{f}_T := T^{-1} \sum_{t=1}^{T} f(a_t, b_t)$.

We write

$$\begin{aligned}
\operatorname{dist}(\bar{f}_t, \mathcal{X}) &= \left\| \bar{f}_t - \tilde{f}_t + \tilde{f}_t - \tilde{\pi}_t + \tilde{\pi}_t - \pi_t \right\|_2 \\
&\leq \left\| \bar{f}_t - \tilde{f}_t \right\|_2 + \left\| \tilde{f}_t - \tilde{\pi}_t \right\|_2 + \left\| \tilde{\pi}_t - \pi_t \right\|_2 \\
&\leq \frac{M+R}{\sqrt{t}} + 2 \left\| \bar{f}_t - \tilde{f}_t \right\|_2 ,
\end{aligned}$$

where we have used the non-expansiveness of projections.

Define $X_t := \sum_{\tau=1}^{t} f(a_\tau, b_\tau) - f(p_\tau, b_\tau)$. Then $(X_t)_{t \in \mathbb{N}}$ is a vector-valued martingale, and $\|X_{t+1} - X_t\|_2 \leq 2M$. By the Azuma inequality, we have

$$\mathsf{P}\left(\left\| \bar{f}_t - \tilde{f}_t \right\|_2 \geq u \right) \leq \mathrm{e}^{2.5} \exp\left(\frac{-ctu^2}{4M^2} \right).$$

Calibrated Forecasting

ε-Calibrated Binary Probability Forecasting

Consider an ε-covering $\mathcal{Q}_\varepsilon := \{q_1, q_2, \ldots, q_{M_\varepsilon}\}$ of the interval $[0, 1]$.

Binary Probability Forecasting Protocol

For $t = 1, 2, 3, \ldots,$

- FORECASTER announces $p_t \in \mathcal{Q}_\varepsilon$.
- REALITY announces $y_t \in \{0, 1\}$, without knowing p_t.

Definition (ε-calibration [3, 9])

A strategy of FORECASTER is ε-*calibrated*, if for any strategy of REALITY,

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{m=1}^{M_\varepsilon} \sum_{t=1}^{T} \mathbb{I}_{\{p_t = q_m\}} \left| q_m - y_t \right| \leq \varepsilon, \quad \text{a.s.}$$

Equivalent Formulation (1/2)

Binary Probability Forecasting Protocol

For $t = 1, 2, 3, \ldots,$

- FORECASTER announces $p_t \in \mathcal{Q}_\varepsilon$.
- REALITY announces $y_t \in \{0, 1\}$, without knowing p_t.

Formulation as a Repeated Game with Vector Payoffs

Set $\mathcal{A} := \mathcal{Q}_\varepsilon$ and $\mathcal{B} := \{0, 1\}$. Define the payoff function

$$f(a, b) := (0, \ldots, 0, a - b, 0, \ldots, 0) \in \mathbb{R}^{M_\varepsilon}.$$

For $t = 1, 2, 3, \ldots,$

- ALICE announces $a_t \in \mathcal{Q}_\varepsilon$.
- BOB announces $b_t \in \{0, 1\}$, without knowing a_t.
- ALICE observes the payoff $f(a, b) \in \mathbb{R}^{M_\varepsilon}$.

Equivalent Formulation (2/2)

Definition (ε-calibration [3, 9])

A strategy of FORECASTER is ε-*calibrated*, if for any strategy of REALITY,

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{m=1}^{M_\varepsilon} \sum_{t=1}^{T} \mathbb{I}_{\{p_t = q_m\}} \left| q_m - y_t \right| \leq \varepsilon, \quad \text{a.s.}$$

Equivalent definition

A strategy of ALICE is ε-calibrated, if $\mathcal{B}_\varepsilon$, the ℓ_1-norm ball of ℓ_1 radius ε, is approachable.

Checking Approachability

Proposition

The ℓ_1-norm ball $\mathcal{B}_\varepsilon$ is approachable by ALICE/FORECASTER.

Proof.

For any $p_\mathsf{B} \in \Delta(\mathcal{B})$, let p_A be the point closest to p_B in $\mathcal{Q}_\varepsilon$; then we have $f(p_\mathsf{A}, p_\mathsf{B}) \in \mathcal{B}_\varepsilon$, verifying the Blackwell condition. □

Recall the achieving strategy for ALICE/FORECASTER.

An ε-Calibrated Forecasting Strategy

Compute

$$P_t \in \operatorname*{arg\,min}_{P \in \Delta(\mathcal{A})} \max_{Q \in \Delta(\mathcal{B})} \left\langle \tilde{f}_t - \mathrm{proj}_{\mathcal{B}_1}(\tilde{f}_t), f(P,Q) - \mathrm{proj}_{\mathcal{B}_1}(\tilde{f}_t) \right\rangle,$$

where $\tilde{f}_T := T^{-1} \sum_{t=1}^{T} f(P_t, b_t)$. Choose $p_t \in \mathcal{Q}_\varepsilon$ randomly, according to P_t.

Non-Asymptotic Result

Theorem

For any $\varepsilon > 0$, there exists an ε-calibrated forecasting strategy, such that for all $T \in \mathbb{N}$, with probability at least $1 - \delta$,

$$\frac{1}{T} \sum_{m=1}^{M_\varepsilon} \sum_{t=1}^{T} \mathbb{I}_{\{p_t = q_m\}} \left| q_m - y_t \right|$$
$$\leq \varepsilon + \sqrt{\frac{2}{\varepsilon}} \left(\sqrt{\frac{1+\varepsilon}{T}} + 4 \sqrt{\frac{1}{cT} \left[\log\left(\frac{1}{\delta}\right) + 2.5 \right]} \right).$$

Proof.

Choose $\mathcal{Q}_\varepsilon$ as the uniform quantization. Then we have $M_\varepsilon \leq 2/\varepsilon$.

$$\frac{1}{T} \sum_{m=1}^{M_\varepsilon} \sum_{t=1}^{T} \mathbb{I}_{\{p_t = q_m\}} \left| q_m - y_t \right|$$
$$= \left\| \bar{f}_t \right\|_1 \leq \left\| \pi_t \right\|_1 + \left\| \bar{f}_t - \pi_t \right\|_1 \leq \left\| \pi_t \right\|_1 + \sqrt{M_\varepsilon} \left\| \bar{f}_t - \pi_t \right\|_2 .$$

□

Existence of a Calibrated Forecasting Strategy

Definition (Calibration [3, 13])

A strategy of FORECASTER is *calibrated*, if for any strategy of REALITY, $\varepsilon > 0$, and $\rho \in [0, 1]$,

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}_{\{|p_t - \rho| \leq \varepsilon\}} \, |p_t - \mathbf{y}_t| = 0, \quad \text{a.s..}$$

Theorem ([13])

There exists a calibrated forecasting strategy.

Proof.

Consider a multiple-round approach. For each round $r \in \mathbb{N}$, set $T_r = 2^r$ and $\varepsilon_r = 2^{-0.5r}$. Check the almost sure convergence by the non-asymptotic result and Borel-Cantelli lemma. □

References I

[1] David Blackwell.
An analog of the minimax theorem for vector payoffs.
Pac. J. Math., 6(1):1–8, 1956.

[2] Nicolò Cesa-Bianchi and Gábor Lugosi.
Prediction, Learning, and Games.
Cambridge Univ. Press, Cambridge, UK, 2006.

[3] A. P. Dawid.
The well-calibrated Bayesian.
J. Am. Stat. Assoc., 77(379):605–610, 1982.

[4] A. P. Dawid.
The impossibility of inductive inference.
J. Am. Stat. Assoc., 80(390):340–341, 1985.

[5] J. L. Doob.
Stochastic Processes.
John Wiley & Sons, New York, NY, 1953.

[6] Dean P. Foster.
A proof of calibration via Blackwell's approachability theorem.
Games Econ. Behav., 29:73–78, 1999.

References II

[7] Dean P. Foster and Sergiu Hart.
Smooth calibration, leaky forecasts, finite recall, and Nash dynamics.
2017.

[8] Dean P. Foster and Rakesh V. Vohra.
Calibrated learning and correlated equilibrium.
Games Econ. Behav., 21:40–55, 1997.

[9] Dean P. Foster and Rakesh V. Vohra.
Asymptotic calibration.
Biometrika, 85(2):379–390, 1998.

[10] Elad Hazan and Sham Kakade.
(weak) calibration is computationally hard.
In *Proc. 25th Annu. Conf. Learning Theory*, 2012.

[11] Anatoli Juditsky and Arkadi Nemirovski.
Solving variational inequalities with monotone operators on domains given by linear minimization oracles.
Math. Program., Ser. A, 156:221–256, 2016.

[12] Sham M. Kakade and Dean P. Foster.
Deterministic calibration and Nash equilibrium.
J. Comput. Syst. Sci., 74:115–130, 2008.

References III

[13] Shie Mannor and Gilles Stoltz.
A geometric proof of calibration.
Math. Oper. Res., 35(4):721–727, 2010.

[14] Assaf Naor.
On the Banach-space-valued Azuma inequality and small-set isoperimetry of Alon-Roichman graphs.
Combinatorics Probab. Comput., 21:623–634, 2012.

[15] Yurii Nesterov.
Barrier subgradient method.
Math. Program., Ser. B, 127:31–56, 2011.

[16] David Oakes.
Self-calibrating priors do not exist.
J. Am. Stat. Assoc., 80(390):339, 1985.

[17] Alvaro Sandroni.
The reproducible properties of correct forecasts.
Int. J. Game Theory, 32:151–159, 2003.

[18] J. v. Neumann.
Zur theorie der gesellschaftsspiele.
Math. Ann., 100:295–320, 1928.

References IV

[19] Vladimir Vovk.
Predictions as statements and decisions.
2006.
arXiv:cs/0606093v1.

[20] Vladimir Vovk.
Non-asymptotic calibration and resolution.
Theor. Comput. Sci., 387:77–89, 2007.