Information Design Perspective on Calibration

PART I – INTRODUCTION

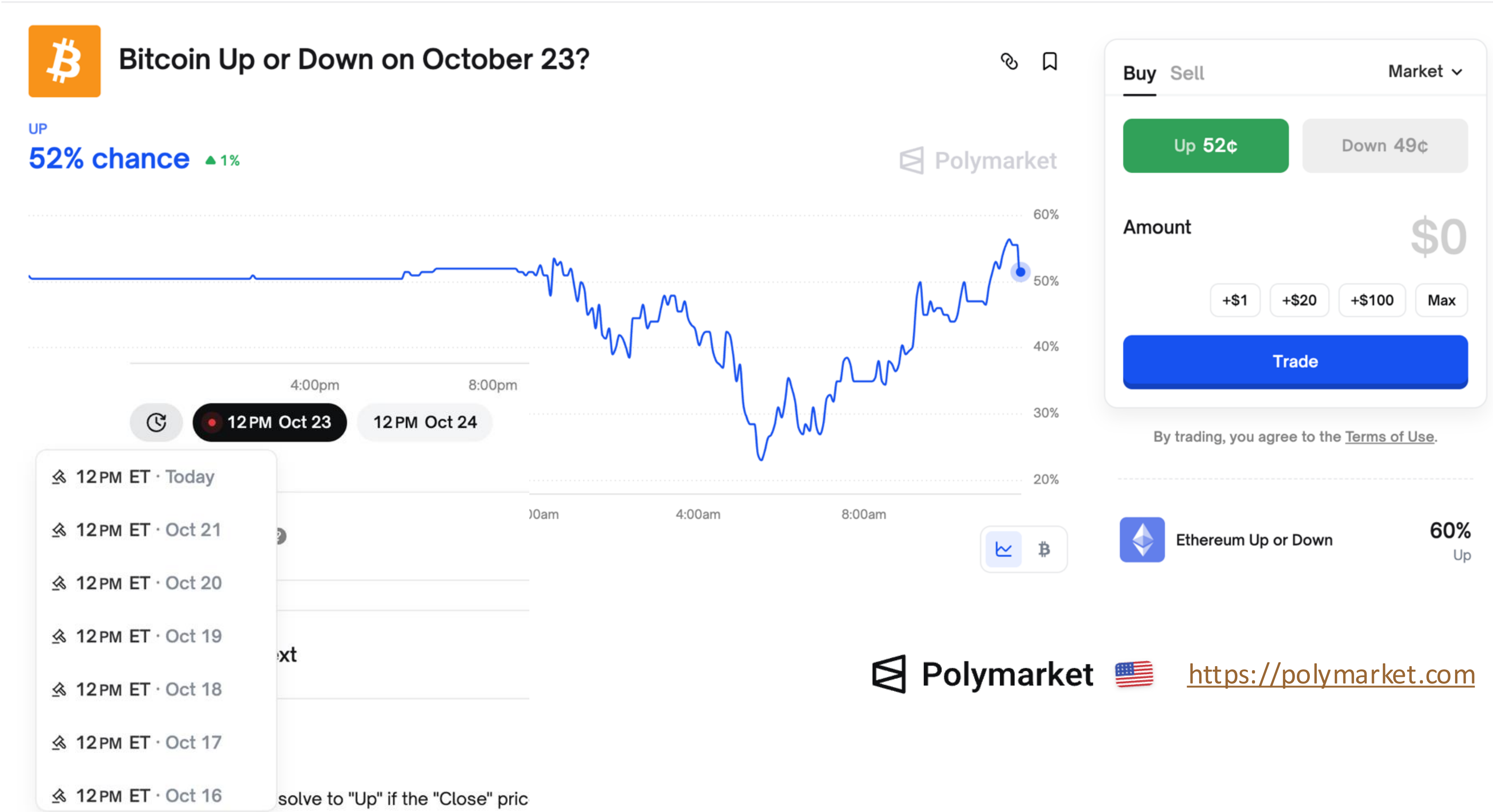
YIDING FENG, HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

JOINT TUTORIAL WITH WEI TANG, CUHK

WINE 2025

Calibration 101: What is **calibration**?

Example: Bitcoin Up or Down?



Polymarket https://polymarket.com

Question: How to measure the **quality** of **prediction**?

Example: Bitcoin Up or Down?

Predictions in past 100 days:

	day 1	day 2	day 3	day 4	…	day 99	day 100
Prediction (prob for **Up**)	50%	20%	20%	50%		70%	20%
Outcome	**Up**	**Up**	**Down**	**Down**		**Down**	**Up**

Natural criterion of **Good** predictions:

- "among all days with prediction = X%, X% of those days are **UP**"

Calibrated Predictor

- Binary random outcome $Y \in \{0, 1\}$ "**Down** vs. **Up**"
- (Possibly random) predictor $F \in \Delta([0, 1])$ "$p \in [0,1]$: chance for **Up**"

Definition [Dawid, JASA'82][Foster Vohra, Biometrika'98].

Predictor F is **calibrated** if for **every** prediction $q \in [0,1]$

$$\mathbb{E}[Y|F = q] = q$$

"condition on prediction, true **Up** probability is equal to prediction"

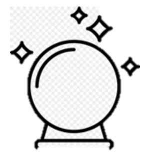
(Mis-)Calibrated Predictors

Example. Suppose $Y \sim \mathrm{Bern}(0.5)$.

calibrated

Predict base rate $\mathbb{E}[Y]$

$$F \equiv \mathbb{E}[Y] = 0.5$$



Predict 100% (resp. 0%) if **Up** (resp. **Down**)

$$F \equiv \mathbb{I}\{Y = 1\}$$

mis-calibrated

Predict 100% and 0% uniformly

$$F \sim \mathrm{Bern}(0.5)$$

F independent of Y

$$\mathbb{E}[Y|F = 1] = 0.5$$
$$\mathbb{E}[Y|F = 0] = 0.5$$

Predict 100% (resp. 0%) if **Down** (resp. **Up**)

$$F \equiv \mathbb{I}\{Y = 0\}$$

$$\mathbb{E}[Y|F = 1] = 0$$
$$\mathbb{E}[Y|F = 0] = 1$$

Calibration Error

Definition

The expected calibration error (ECE) of predictor F is

$$\mathrm{ECE}[F] \stackrel{\text{def}}{=} \mathbb{E}_{q \sim F}[|q - \mathbb{E}[Y|q]|]$$

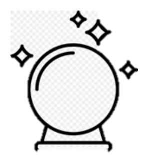
We say a predictor F is ε-calibrated if $\mathrm{ECE}[F] \leq \varepsilon$.

Example (cont.) Suppose $Y \sim \mathrm{Bern}(0.5)$. Compute ECE:

Predict base rate $\mathbb{E}[Y]$

$$F \equiv \mathbb{E}[Y] = 0.5$$

$$\mathrm{ECE}[F] = |0.5 - 0.5| = 0$$



Predict 100% (resp. 0%) if **Up** (resp. **Down**)

$$F \equiv \mathbb{I}\{Y = 1\}$$

$$\mathrm{ECE}[F] = |1 - 1| \cdot 0.5 + |0 - 0| \cdot 0.5 = 0$$

Predict 100% and 0% uniformly

$$F \sim \mathrm{Bern}(0.5)$$

F independent of Y

$$\mathrm{ECE}[F] = |0.5 - 1| \cdot 0.5 + |0.5 - 0| \cdot 0.5 = 0.5$$

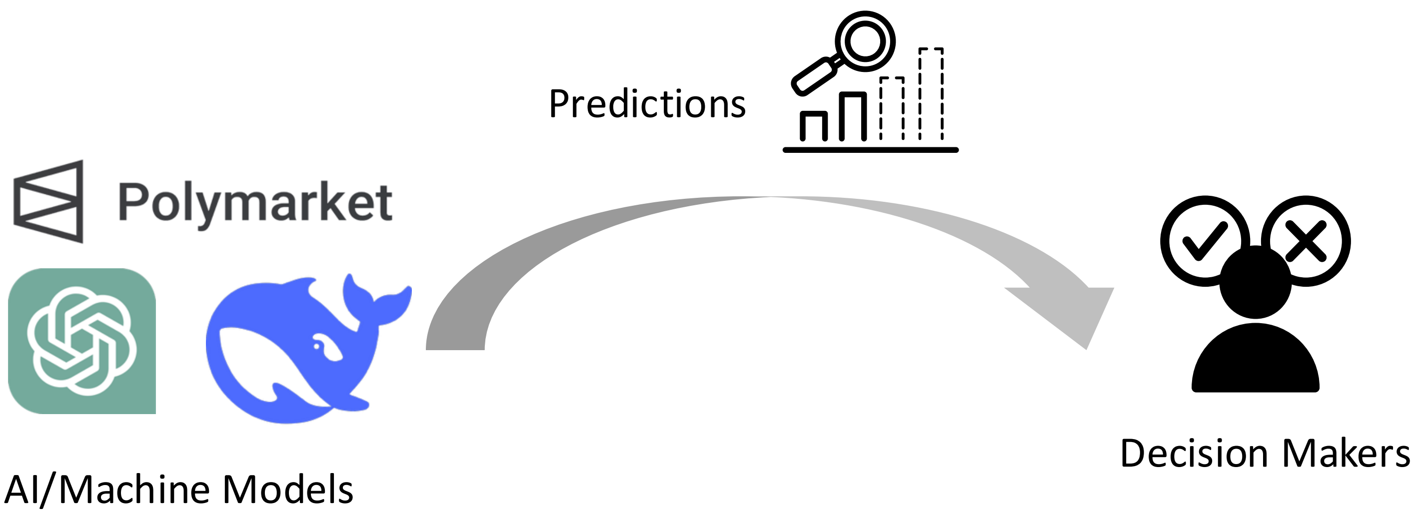
Predict 100% (resp. 0%) if **Down** (resp. **Up**)

$$F \equiv \mathbb{I}\{Y = 0\}$$

$$\mathrm{ECE}[F] = |0 - 1| \cdot 0.5 + |1 - 0| \cdot 0.5 = 1$$

Why **calibration**?

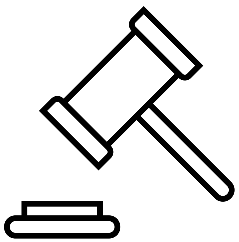
Machine-in-the-loop Decision Making



Other applications of machine predictors:

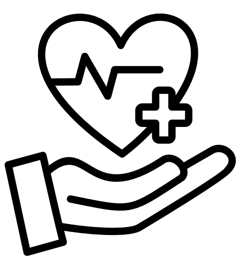
- Digital advertising
- Criminal Justice
- Healthcare
- …

Machine Predictions in High-Stakes Decision Making



- **Digital Advertising Auction: Google's CTR models**
  - offer CTR prediction for advertisers to guide their auction bidding

- **Criminal Justice: COMPAS (Northpointe/Equivant)**
  - offer recidivism predictions to guide judges on bail decisions



- **Healthcare**: **Epic's Deterioration Index / Sepsis Model**
  - provides disease predictions to guide doctor's clinic decision

Trustworthiness in Machine Learning

"**calibrated** predictor $\Longrightarrow$ **reliable** to be used to make decisions"

Downstream decision maker (agent)

- Unobserved binary random outcome $Y \in \{0, 1\}$
- Decides action $a \sim \mathcal{A}$
- Receives agent utility $u(a, Y) \in [0,1]$

"**Up** vs. **Down**"

"**Long** vs **Short** Bitcoin?"

Informal Theorem [Kleinberg Leme Schneider Teng, COLT'23]**.**

Given any predictor F, suppose agent **naively best respond** to prediction $q \sim F$, i.e.,

$$a^* = \max_{a \in \mathcal{A}} \mathbb{E}_{Y \sim \mathrm{Bern}(q)}[u(a, Y)]$$

then agent's **regret** is at most $\mathrm{ECE}[F]$.

Regret = Payoff by best-responding to $\mathbb{E}[Y \mid q] -$ **Payoff by best-responding to** q

Trustworthiness in Machine Learning

Takeaways

1. **No-regret**: Making decision based on perfectly/almost calibrated predictor ensures zero/small regret.
2. **No need** to know details of F

Related Work

Econ/Stats literature [Foster Vohra, GEB’97, Biometrika’98], [Hart Mas-Colell, ECMA’00], [Foster Hart, JPE’21], [Foster Hart, TE’23], [Guo Shmaya, TE’23] …

CS/ML literature:

- **Calibration in neural network/LLMs** [Guo Pleiss Sun Weinberger, ICML’17], [Percy Liang et al., TMLR’23], [Kalai Vempala, STOC’24], …
- **Online calibration** [Qiao Valiant, STOC’21], [Qiao Zhang, COLT’24], [Okoroafor Kleinberg Sun, AISTATS’24], [Dagan Daskalakis Fishelson Golowich Kleinberg Okoroafor, STOC’25], …
- **Multi-calibration** [Hébert-Johnson Kim Reingold Rothblum, ICML’18], [Hu Peale, ITCS’23], [Casacuberta Dwork Vadhan, STOC’24], …
- **Calibration in Decision-making** [Camara Hartline Johnsen FOCS’20], [Rothblum Yona, ITCS’23], [Kleinberg Leme Schneider Teng, COLT’23], [Roth Shi, EC’24], [Jain Vianney, EC’24], [Collina Roth Shao, EC’24], [Hu Wu, FOCS’24], [Feng Tang, SODA’26] …

OR/MS literature [Elmachtoub Grigas, MS’22], [Jens Witkowski et al, MS’23], [Tsirtsis Tabibian Khajehnejad Singla Scholkopf Gomez-Rodriguez, MS’24] …

Related Work

On calibration of modern neural networks

C Guo, G Pleiss, Y Sun, KQ Weinberger

International conference on machine learning, 2017 • proceedings.mlr.press

Abstract

Confidence calibration–the problem of predicting probability estimates representative of the true correctness likelihood–is important for classification models in many applications. We discover that modern neural networks, unlike those from a decade ago, are poorly calibrated. Through extensive experiments, we observe that depth, width, weight decay, and Batch Normalization are important factors influencing calibration. We evaluate the performance of various post-processing calibration methods on state-of-the-art

SHOW MORE

Save Cite Cited by 8375 Related articles All 8 versions

Information Design 101: Bayesian Persuasion

Core idea: a *sender* commits to an information structure to influence a *receiver*'s action

Basic model [Kamenica Gentzkow, AER'11]

- State $\omega \in \Omega$ drawn from prior μ
- Sender *commits* to signaling scheme $\pi: \Omega \to \Delta(\Sigma)$, where signal $\sigma \in \Sigma$ is revealed to receiver
- Receiver observes σ, updates belief via Bayes' rule, choose action $a \in \mathcal{A}$ to maximize their utility
- Sender chooses π to maximize their own utility, anticipating receiver's best response.

Key insight: Sender cannot lie after seeing ω, but can *design what is revealed* in advance.

Example: ad platform (sender) designs how to report predicted CTR to advertiser (receiver), who then decides how much to bid

Calibration: Information Design Perspective

Calibration

- Binary random outcome $Y \in \{0, 1\}$

- Feature/Context $X_i \sim D \in \Delta(\mathcal{X})$

- Each X_i has true prob. $p_i \in [0,1]$

  for outcome realization:

$$\mathbb{E}[Y|X = X_i] \stackrel{\text{def}}{=} p_i$$

- Predictor $F: \mathcal{X} \to \Delta([0, 1])$

Calibrated predictor requires that:

$$q = \mathbb{E}[Y|q] = \frac{\sum_{X_i} \mathbb{P}_{X\sim D}(X = X_i) \cdot F(q \mid X_i) \cdot p_i}{\sum_{X_i} \mathbb{P}_{X\sim D}(X = X_i) \cdot F(q \mid X_i)}, \qquad \forall q \in \text{supp}(F)$$

Calibration: Information Design Perspective

Calibration

- Binary random outcome $Y \in \{0, 1\}$
- Feature/Context $X_i \sim D \in \Delta(\mathcal{X})$
- Each X_i has true prob. $p_i \in [0,1]$ for outcome realization:

$$\mathbb{E}[Y|X = X_i] \stackrel{\text{def}}{=} p_i$$

- Predictor $F: \mathcal{X} \to \Delta([0, 1])$

Signaling scheme

- State space $\{p_i\}$
- Prior dist. is $\mathbb{P}(p = p_i) = \mathbb{P}_{X\sim D}(X = X_i)$
- Signal Space $\Sigma = [0, 1]$, signaling scheme $\pi: \{p_i\} \to \Delta(\Sigma)$

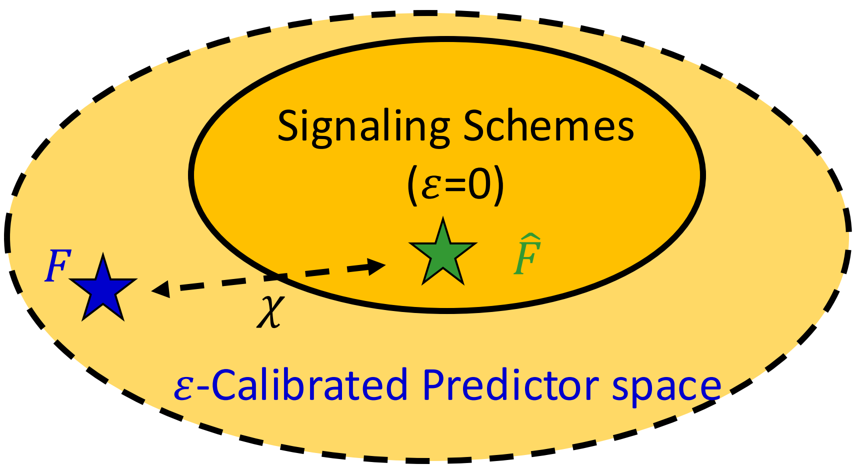
Calibrated predictor ⇔ signaling scheme where signal is posterior mean

$$q = \mathbb{E}[Y|q] = \frac{\sum_{X_i} \mathbb{P}_{X\sim D}(X = X_i) \cdot F(q \mid X_i) \cdot p_i}{\sum_{X_i} \mathbb{P}_{X\sim D}(X = X_i) \cdot F(q \mid X_i)}, \qquad \forall q \in \mathrm{supp}(F)$$

Two-Step View for Predictors

Observation: generating ε-calibrated predictor F is equivalent to

1. generating **(perfectly) calibrated predictor** $\hat{F}$ (ECE $\varepsilon = 0$)
   - **perfectly calibrated predictor** $\Leftrightarrow$ signaling scheme where **signals** $=$ **posterior means**
2. "miscalibrating" $\hat{F}$ into F with ε-calibration budget (denote this **miscalibration as** χ)



Examples via Two-Step Approach

Define $\chi(q, q')$: frequency of miscalibrating true prob. q to prediction q'

calibrated predictor

Predict base rate $\mathbb{E}[Y]$

$$\hat{F} \equiv \mathbb{E}[Y] = 0.5$$

$$\chi(0.5, 0) = 0.5$$

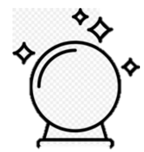
$$\chi(0.5, 1) = 0.5$$

miscalibrated predictor

Predict 100% and 0% uniformly

$$F \sim \mathrm{Bern}(0.5)$$

F independent of Y



Predict 100% (0%) if **Up** (**Down**)

$$\hat{F} \equiv \mathbb{I}\{Y = 1\}$$

$$\chi(1, 0) = 1$$

$$\chi(0, 1) = 1$$

Predict 100% (0%) if **Down** (**Up**)

$$F \equiv \mathbb{I}\{Y = 0\}$$

Two-Step View for Predictors

Space of ϵ-calibrated predictor can be characterized as linear polytope:

Variable $\hat{F}$: perfectly calibrated predictor

Variable $\chi(q, q')$: frequency of miscalibrating true prob. q to prediction q'

Linear constraints:

$$\sum_{q\in[0,1]} \sum_{q'\in[0,1]} \chi(q, q') \cdot |q' - q| \leq \varepsilon$$

χ satisfies ε-ECE budget

similar to budget constraint in auction design

$$\sum_{q'\in[0,1]} \chi(q, q') = \widehat{F}(q), \forall q$$

χ is consistent with $\hat{F}$

$\hat{F}$ is perfectly calibrated

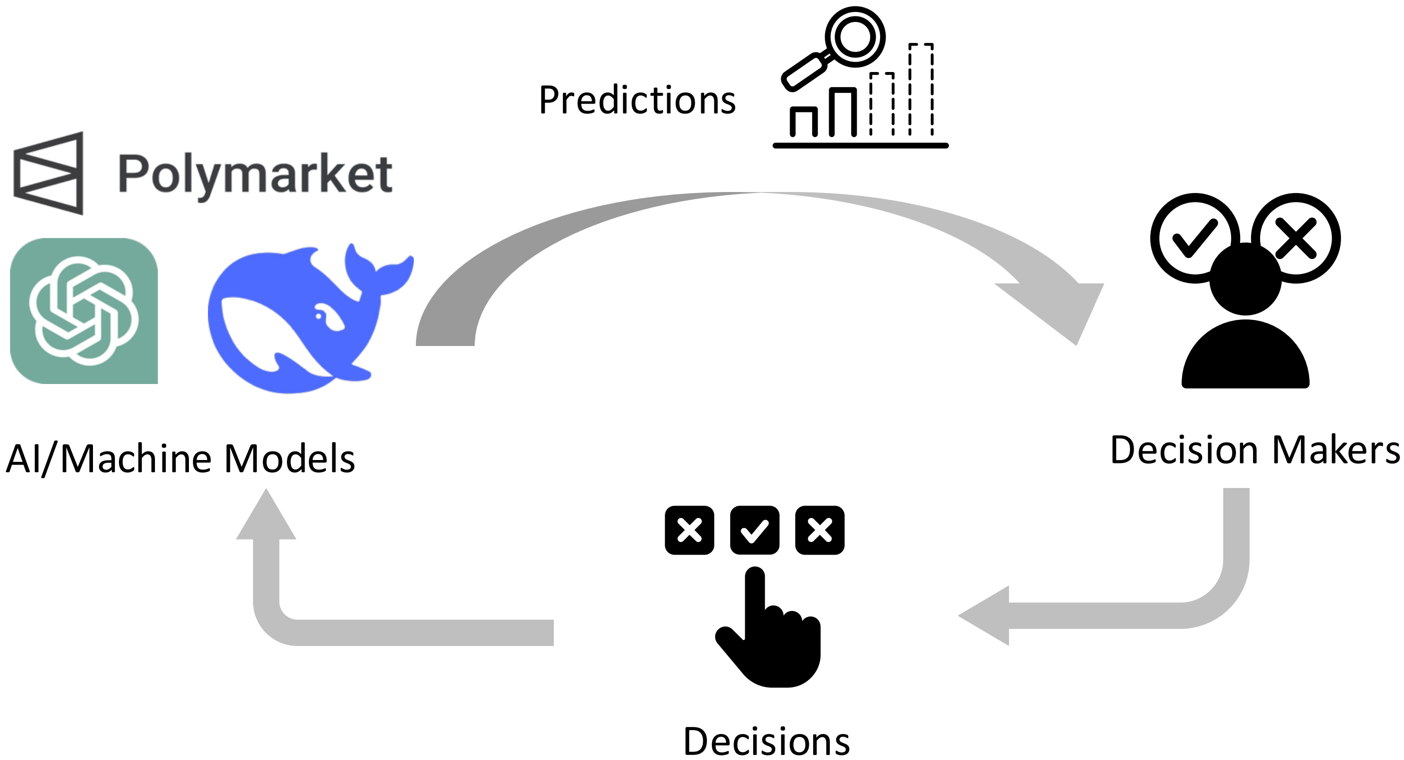
Captured by **mean-preserving contractions (MPC) constraint**

i.e., 2nd-order stochastic dominance + identical mean

widely studied in information design literature

Takeaway: Useful tool for characterizing/computing (near)-optimal predictors

Application: Predictor Design under Incentive Misalignment



Decision maker’s action may also affect AI designer’s utility

Application: Predictor Design under Incentive Misalignment

[Informal Research Questions]

[Q1] Given a **calibration error budget,** what is the **optimal predictor**, especially when there exists **incentive misalignment** between the principal and the agent?

[Q2] Can we **compute** this optimal predictor or an approximately optimal predictor in **an efficient way**?

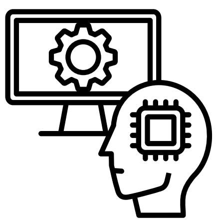
Persuasive Calibration

- Binary random outcome $Y \in \{0, 1\}$
- Feature/Context X sampled from feature dist. $D \in \Delta(\mathcal{X})$
- Each feature X has true prob. $p_X \in [0,1]$ for outcome realization:

$$\mathbb{E}[Y|X] \stackrel{\text{def}}{=} p_X$$

Principal

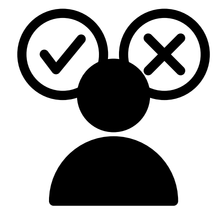
(machine designer)



- Knows D, p_X, but not Y
- Principal utility: $u^P(a)$

Agent

(decision maker)



- Decides action $a \sim \mathcal{A}$
- Agent utility: $u^A(a, Y)$

Persuasive Calibration

random mapping from feature to predictions

Goal: identify principal's optimal predictor $F\colon \mathcal{X} \to \Delta([0,1])$ subject to ECE constraint

$$\max_{F:\mathcal{X}\to\Delta([0,1])} \quad u^P(F) = \mathbb{E}_{X\sim D}\mathbb{E}_{q\sim F(X)}[u^P(a^*(q))]$$

$$s.t. \quad \mathrm{ECE}[F] \le \varepsilon$$

ECE constraint, ε is pre-specified **ECE budget**

$$a^*(q) = \boxed{\operatorname*{argmax}_{a\in\mathcal{A}}} \mathbb{E}_{\boxed{Y\sim\mathrm{Bern}(q)}}[u^A(a,Y)], \quad \boxed{\forall q \in \mathrm{supp}(F)}$$

agent **naively best responds**

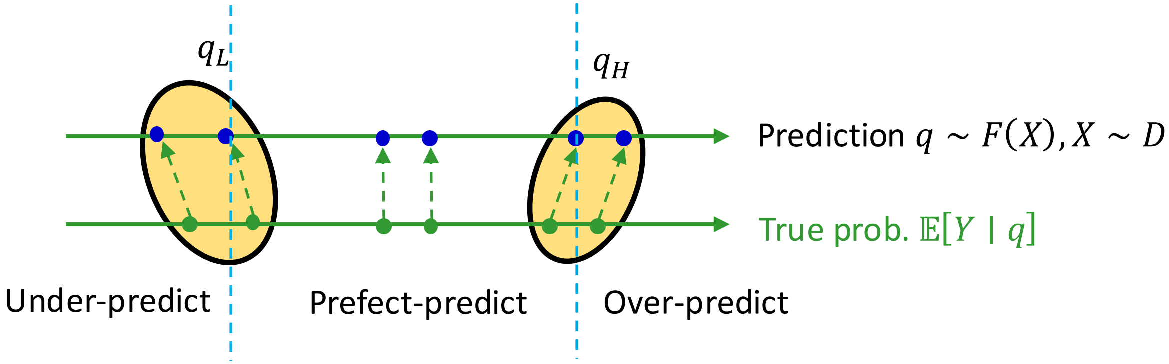
trustworthiness ensured by ε-ECE above

Bayesian Persuasion	**Persuasive Calibration**
Knowing Priors & All signaling details	

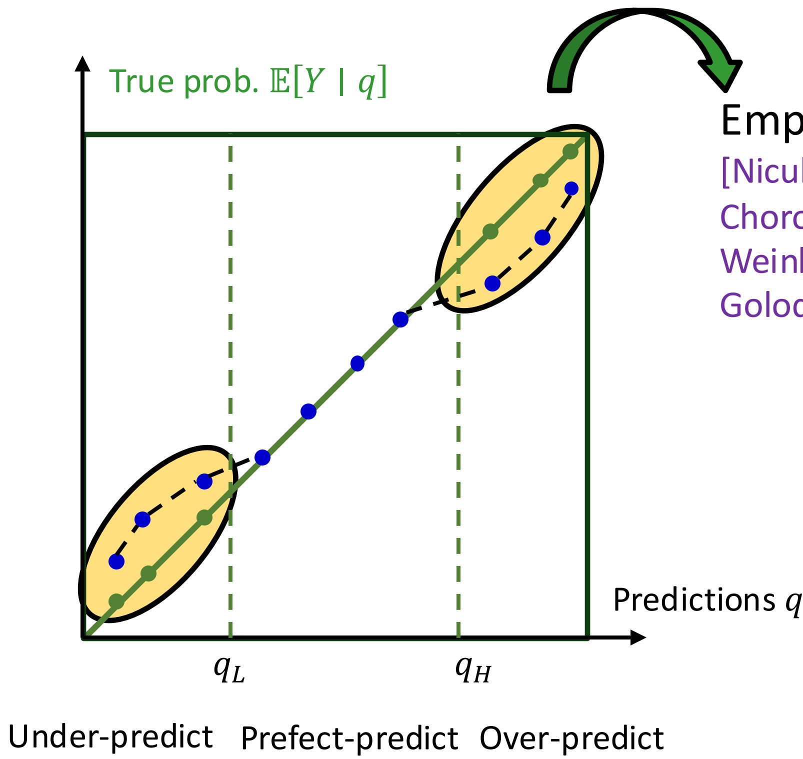
Characterizing Optimal ε-Calibrated Predictor

Theorem [Feng Tang, SODA'26]**.** In the optimal ε-calibrated predictor, there exists $0 \le q_L \le q_H \le 1$ such that

- [Miscalibration structure] Predictions $q \ge q_H$ ($q \le q_L$) over-predict (under-predict) true conditional probability.



Over-/under-confident Predictions in ML



Empirical evidence:

[Niculescu-Mizil Caruana, ICML'05] [Pereyra Tucker Chorowski Kaiser Hinton, ICLR'17] [Guo Pleiss Sun Weinberger, ICML'17] [Mukhoti Kulharia Sanyal Golodetz Torr Dokania, NeurIPS'20] …

Takeaways

Regardless of loss functions, as long as there is calibration error, it must happen on extreme predictions.

Characterizing Optimal ε-Calibrated Predictor

Theorem [Feng Tang, SODA'26]**.** In the optimal ε-calibrated predictor, there exists $0 \le q_L \le q_H \le 1$ such that

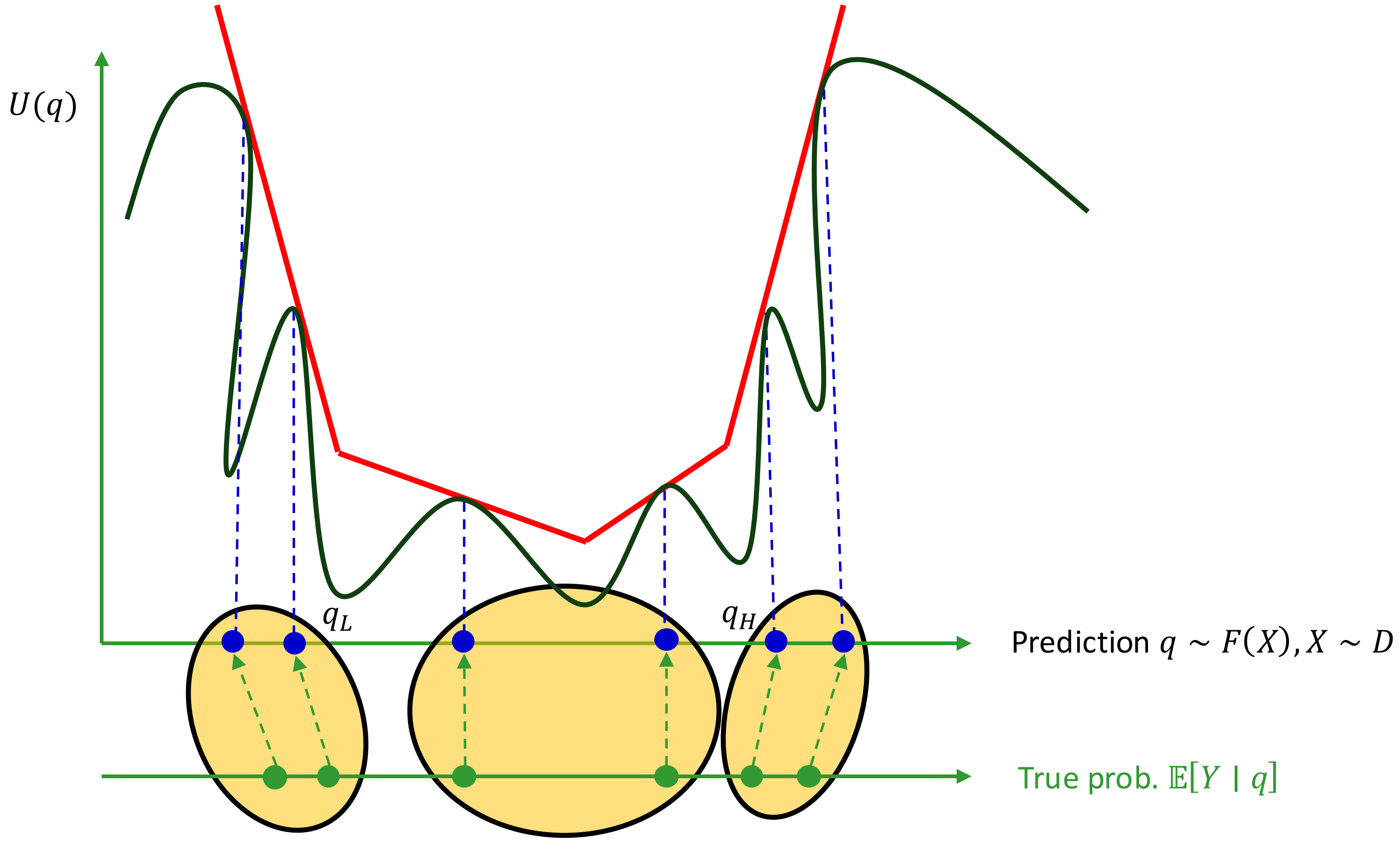
- [Miscalibration structure] Predictions $q \ge q_H$ ($q \le q_L$) over-predict (under-predict) true conditional probability.
- [Payoff structure] For all predictions $q \in \operatorname{supp}(F)$, the derivative of principal's **indirect utility function** $U'(q)$ is **increasing** in q, and satisfies

$$U'(q) = \alpha \text{ for } q \ge q_H;\ U'(q) = -\alpha \text{ for } q \le q_L$$

$U'(q)$: "**marginal utility gain** by miscalibrating q"

Indirect utility $U(q) \stackrel{\text{def}}{=} u^P\big(a^*(q)\big)$ with $a^*(q) \stackrel{\text{def}}{=} \underset{a \in \mathcal{A}}{\operatorname{argmax}}\ \mathbb{E}_{Y\sim\mathrm{Bern}(q)}[u^A(a, Y)]$

Payoff Structure of Optimal ε-Calibrated Predictor



Proof Sketch via Two-Step LP

Two-step view **linear program** for principal's problem:

Variable $\hat{F}$: perfectly calibrated predictor

Variable $\chi(q, q')$: frequency of miscalibrating true prob. q to prediction q'

$$\max_{\hat{F},\chi} \quad U^P(\chi) := \sum_{q\in[0,1]} \sum_{q'\in[0,1]} \chi(q, q') \cdot U^P(q')$$

$$s.t. \quad \sum_{q\in[0,1]} \sum_{q'\in[0,1]} \chi(q, q') \cdot |q' - q| \leq \varepsilon$$

χ satisfies ε-ECE budget

similar to budget constraint in auction design

$$\sum_{q'\in[0,1]} \chi(q, q') = \hat{F}(q), \forall q$$

χ is consistent with $\hat{F}$

$\hat{F}$ is perfectly calibrated

Captured by **mean-preserving contractions (MPC) constraint**

i.e., 2nd-order stochastic dominance + identical mean

widely studied in information design literature

Proof Sketch via Two-Step LP

Proof ideas for optimal structure:

[three-interval **miscalibration structure**] $\Longleftarrow$ mean-preserving contraction (MPC) constraint

Suppose structure is violated, we can construct another perfectly calibrated predictor $\hat{F}^*$ and miscalibration χ^* with desired miscalibration structure and

- objective is the same
- smaller calibration error

[three-interval **payoff structure**] $\Longleftarrow$ MPC constraint + ECE budget constraint for χ

Proved by LP duality. The analysis shares similarity to auction design for budgeted buyers

- **monotone** $U'(q) \approx$ **monotone allocation rule** in auction design
- **linear tail** $U'(q) = -\alpha$ for $q \leq q_L$, $U'(q) = \alpha$ for $q \geq q_H$ α: dual variable for ECE budget constraint

Computing (Near)-Optimal Predictor

Theorem [Feng Tang, SODA'26] There exists LP-based algorithm for computing optimal ε-calibrated predictor with running time $\mathrm{poly}(|\mathcal{X}|, |\mathcal{A}|)$.

$\mathcal{X}$: feature space, $\mathcal{A}$: action space

- Two-step LP + a novel two-layer discretization $\Longrightarrow$ FPTAS

  apply to more general ℓ_p-ECE constraints

- **Observation**: when $\varepsilon = 0$, **persuasive calibration** $\equiv$ **Bayesian persuasion (BP)**

  OPT can be efficiently computed by applying **revelation principle** and then consider LP of **incentive compatible (IC)** action recommendation

- *Proof idea*: when $\varepsilon > 0$, persuasive calibration can be interpreted a new variant of BP: **persuasion with signal-dependent bias**

  "aggregate IC violation can be at most ε"

Summary

- An intrinsic connection between **calibration** and **information design**
  - Calibrated predictor ≡ signaling scheme where signal = posterior mean
  - [Two-step view] General predictor ≡ calibrated predictor + miscalibration plan
- Applications:
  - Persuasive calibration: how to **design** predictors under incentive misalignment
  - How to **compare** different predictors (next part)
  - How to **design** predictors in digital advertising auction (next part)

Thanks!

Questions?

Please send me an email for any questions/comments:
ydfeng@ust.hk
wtang2359@gmail.com