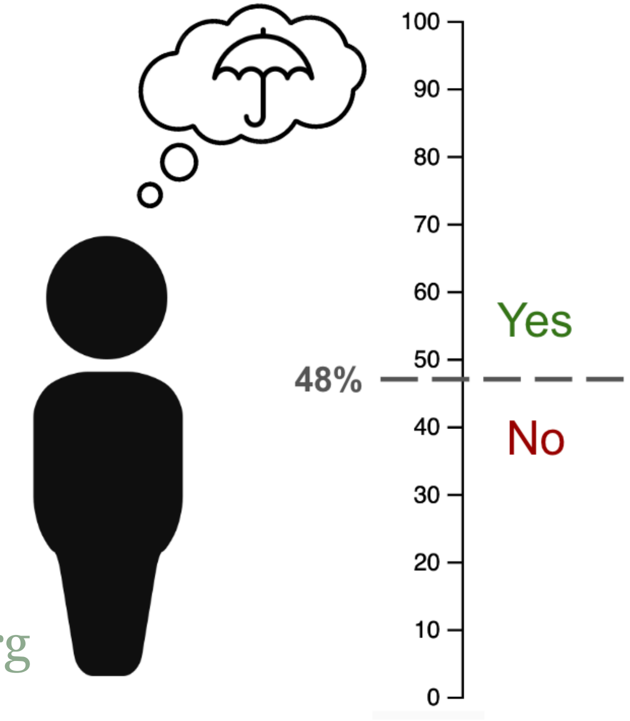
Calibration and Trustworthy Decision Making

Princewill Okoroafor

Cornell University

(Incoming Postdoc at Harvard SEAS)



Based on joint work with Michael P. Kim & Robert Kleinberg

https://arxiv.org/abs/2501.17205

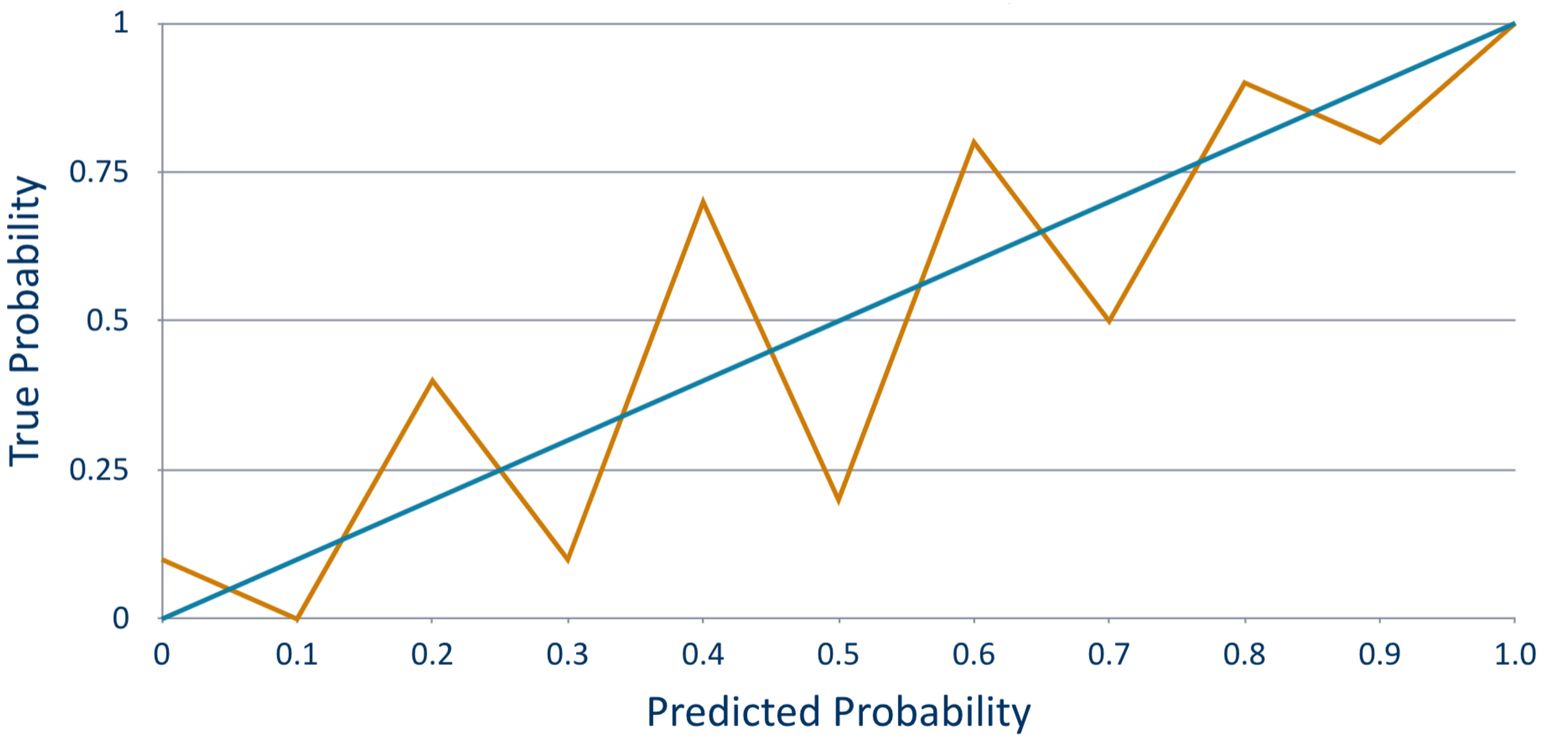
What is calibration?

Prediction values should mean what they say

Sun 07	51°/30°	AM Clouds/PM Sun	14%
Mon 08	61°/41°	Partly Cloudy	7%
Tue 09	69°/47°	Partly Cloudy	21%
Wed 10	59°/47°	Showers	56%
Thu 11	62°/52°	Showers	70%
Fri 12	56°/41°	Showers	58%
Sat 13	52°/39°	AM Showers	32%

What is calibration?

Prediction values should mean what they say



Sun 07	51°/30°	AM Clouds/PM Sun	14%
Mon 08	61°/41°	Partly Cloudy	7%
Tue 09	69°/47°	Partly Cloudy	21%
Wed 10	59°/47°	Showers	56%
Thu 11	62°/52°	Showers	70%
Fri 12	56°/41°	Showers	58%
Sat 13	52°/39°	AM Showers	32%

Binary Prediction Setting

Applications in weather forecasting, prediction markets, etc

Each day $t = 1,2,\ldots,T$,

- Nature chooses an outcome $y_t \in \{0,1\}$,
- Before observing outcome, Forecaster makes a prediction $p_t \in [0,1]$

Goal: low calibration error

Binary Prediction Setting

Applications in weather forecasting, prediction markets, etc

Each day $t = 1,2,\ldots,T$,

- Nature chooses an outcome $y_t \in \{0,1\}$,
- Before observing outcome, Forecaster makes a prediction $p_t \in [0,1]$

Goal: low calibration error

sum of outcomes when p was predicted

$$\mathrm{calerr}(t) = \sum_{p\in[0,1]} n_t(p) \left| p - \underbrace{\frac{m_t(p)}{n_t(p)}}_{\text{average of outcomes when p was predicted}} \right|$$

number of times p was predicted

Calibration is very non-smooth

T =	1	2	…	T/2	…	T
Nature	0	0	0	1	1	1
Forecaster	$\frac{1}{2}-\epsilon$	$\frac{1}{2}-\epsilon$	$\frac{1}{2}-\epsilon$	$\frac{1}{2}+\epsilon$	$\frac{1}{2}+\epsilon$	$\frac{1}{2}+\epsilon$

Calibration is very non-smooth

T =	1	2	…	T/2	…	T
Nature	0	0	0	1	1	1
Forecaster	$\frac{1}{2}-\epsilon$	$\frac{1}{2}-\epsilon$	$\frac{1}{2}-\epsilon$	$\frac{1}{2}+\epsilon$	$\frac{1}{2}+\epsilon$	$\frac{1}{2}+\epsilon$

Smooth Calibration/Distance to Calibration [KF'08, FH'18, BGHN'23]

Calibration is not incentive-compatible

T =	1	…	T/m	…	2T/m	…	T
Nature	$\text{Ber}\left(\frac{1}{m}\right)$	$\text{Ber}\left(\frac{1}{m}\right)$	$\text{Ber}\left(\frac{2}{m}\right)$	$\text{Ber}\left(\frac{2}{m}\right)$	$\text{Ber}\left(\frac{3}{m}\right)$	…	…
Forecaster	$\frac{1}{m}$	$\frac{1}{m}$	$\frac{2}{m}$	$\frac{2}{m}$	$\frac{3}{m}$	…	…

Calibration is not incentive-compatible

T =	1	…	T/m	…	2T/m	…	T
Nature	$\mathrm{Ber}\left(\frac{1}{m}\right)$	$\mathrm{Ber}\left(\frac{1}{m}\right)$	$\mathrm{Ber}\left(\frac{2}{m}\right)$	$\mathrm{Ber}\left(\frac{2}{m}\right)$	$\mathrm{Ber}\left(\frac{3}{m}\right)$	…	…
Forecaster	$\frac{1}{m}$	$\frac{1}{m}$	$\frac{2}{m}$	$\frac{2}{m}$	$\frac{3}{m}$	…	…

Proper Scoring Losses e.g square loss are incentive-compatible

Why Calibrate?

Key Question:

If full calibration has all these flaws that can be addressed by weaker notions, why bother with full calibration?

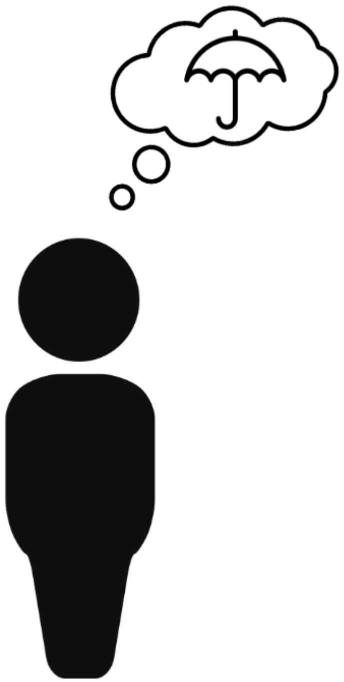
Why Calibrate?

Key Question:

If full calibration has all these flaws that can be addressed by weaker notions, why bother with full calibration?

Calibrated forecasts implies low regret decision making

Sequential Decision Making



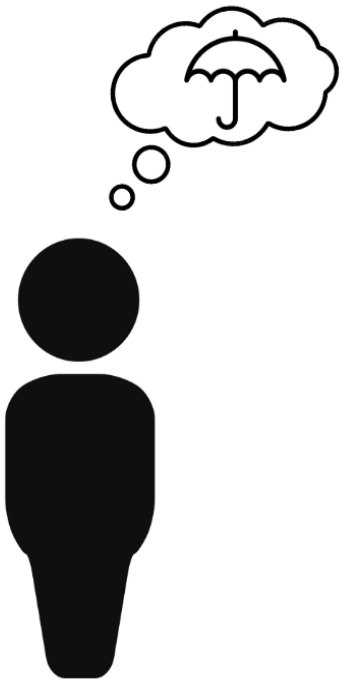
Each day $t = 1, 2, \ldots, T$,

- Nature chooses an outcome $y_t \in \{0,1\}$,
- Decision Maker chooses action $a_t \in A$
- Decision Maker incurs loss $\ell(a_t, y_t) \in [0,1]$

Goal: Low Regret

$$\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in A} \sum_{t=1}^{T} \ell(a, y) \leq o(T)$$

Sequential Decision Making



Each day $t = 1,2,\ldots,T$,

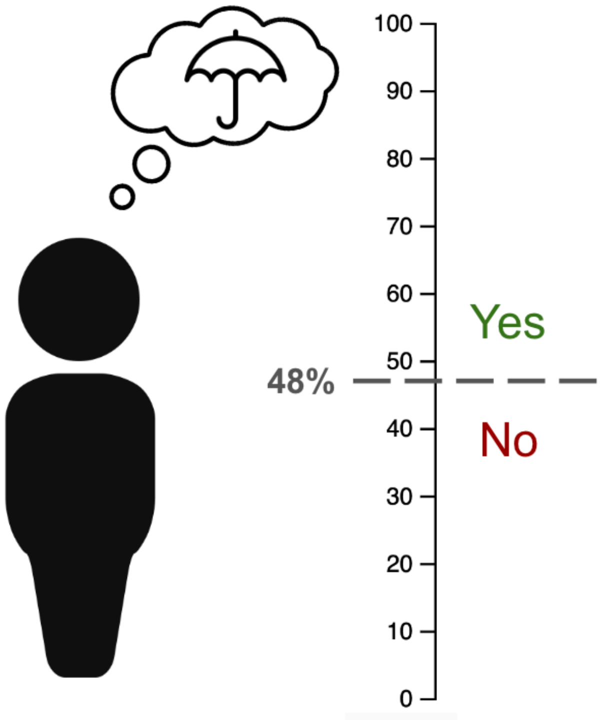
- Nature chooses an outcome $y_t \in \{0,1\}$,
- Decision Maker chooses action $a_t \in A$
- Decision Maker incurs loss $\ell(a_t, y_t) \in [0,1]$

Goal: Low Regret

$$\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in A} \sum_{t=1}^{T} \ell(a, y) \leq o(T)$$

Multiplicative Weights guarantees regret of $O\left(\sqrt{T \log |A|}\right)$

Sequential Decision Making



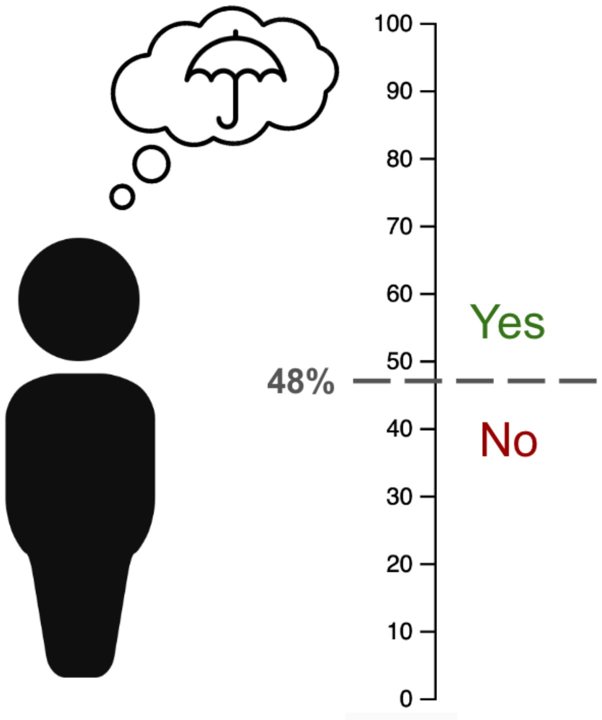
Each day $t = 1,2,\ldots,T$,

- Nature chooses an outcome $y_t \in \{0,1\}$,
- Forecaster makes a prediction $p_t \in [0,1]$
- Decision Maker best-responds to prediction i.e chooses

$$a_t \in \mathrm{argmin}_{a \in A} \mathbb{E}_{y \sim \mathrm{Ber}(p_t)} \left[\ell(a, y) \right]$$

Low Regret: $\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in A} \sum_{t=1}^{T} \ell(a, y) \leq o(T)$

Sequential Decision Making



Each day $t = 1,2,\ldots,T$,

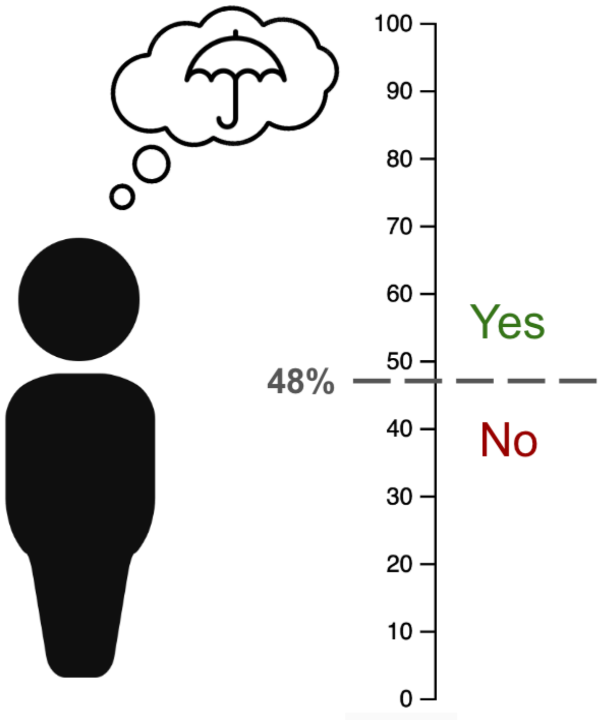
- Nature chooses an outcome $y_t \in \{0,1\}$,
- Forecaster makes a prediction $p_t \in [0,1]$
- Decision Maker best-responds to prediction i.e chooses

$$a_t \in \text{argmin}_{a \in A} \mathbb{E}_{y \sim \text{Ber}(p_t)} \left[\ell(a, y) \right]$$

Low Regret: $\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in A} \sum_{t=1}^{T} \ell(a, y) \leq o(T)$

If predictions are calibrated, then Agent achieves low regret

Sequential Decision Making



Each day $t = 1,2,\ldots,T$,

- Nature chooses an outcome $y_t \in \{0,1\}$,
- Forecaster makes a prediction $p_t \in [0,1]$
- Decision Maker best-responds to prediction i.e chooses

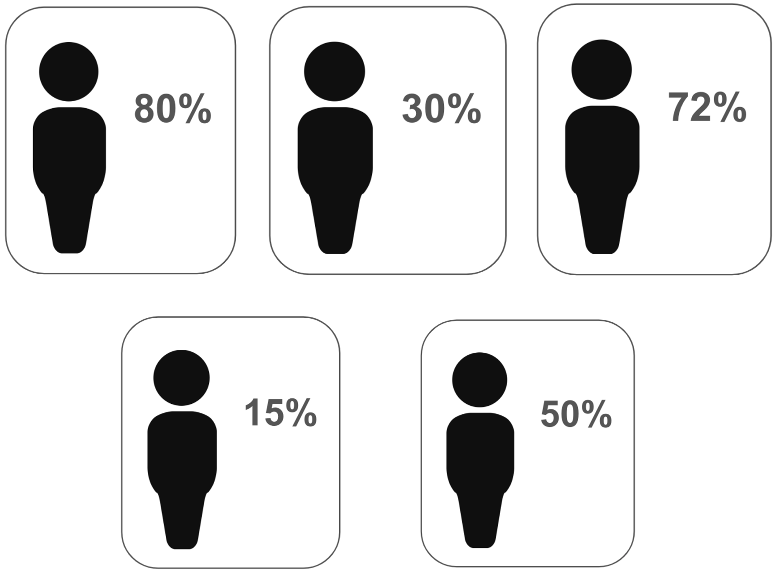
$$a_t \in \operatorname{argmin}_{a \in A} \mathbb{E}_{y \sim \mathrm{Ber}(p_t)} \left[\ell(a, y) \right]$$

Low Regret: $\sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in A} \sum_{t=1}^{T} \ell(a, y) \leq o(T)$

Agent's regret is bounded by Calibration error rate $f(T)$

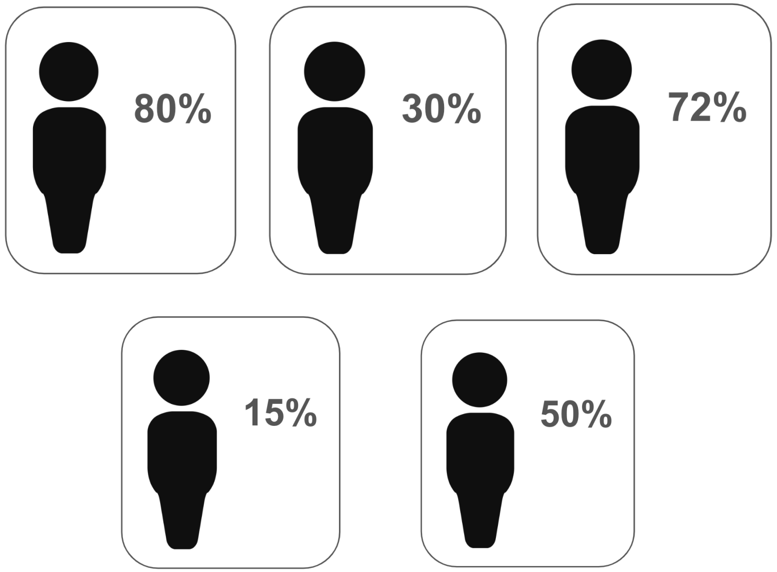
Trustworthy Decision Making

Participants at COLT Workshop



Trustworthy Decision Making

Participants at COLT Workshop



Every agent wants to obtain low regret with respect to their loss function $\ell \in L$

$$\sup_{\ell \in L} \sum_{t=1}^{T} \ell(a_t, y_t) - \min_{a \in A} \sum_{t=1}^{T} \ell(a, y) \leq o(T)$$

Theorem: (KLST'23)

Calibrated forecasts guarantee low regret (at a rate of $f(T)$) simultaneously for every decision maker

KLST'23 - U-Calibration: Forecasting for an unknown agent

The rate of online calibration

Lower Bound: (QV'21, DDFGK**O**'25)

No algorithm can guarantee calibration at a rate better than $\Omega\left(T^{0.543}\right)$

Upper Bound: (FV'98, DDFGK**O**'25)

There exists an algorithm that guarantees calibration at a rate of $O\left(T^{2/3-\epsilon}\right)$

QV'21 - Stronger Lower Bounds for Calibration via Sidestepping
DDFGKO'25 - Breaking the T^2/3 Barrier for Sequential Calibration
FV'98 - Forecast Hedging and Calibration

The rate of online calibration

Lower Bound: (QV'21, DDFGK**O**'25)

No algorithm can guarantee calibration at a rate better than $\Omega\left(T^{0.543}\right)$

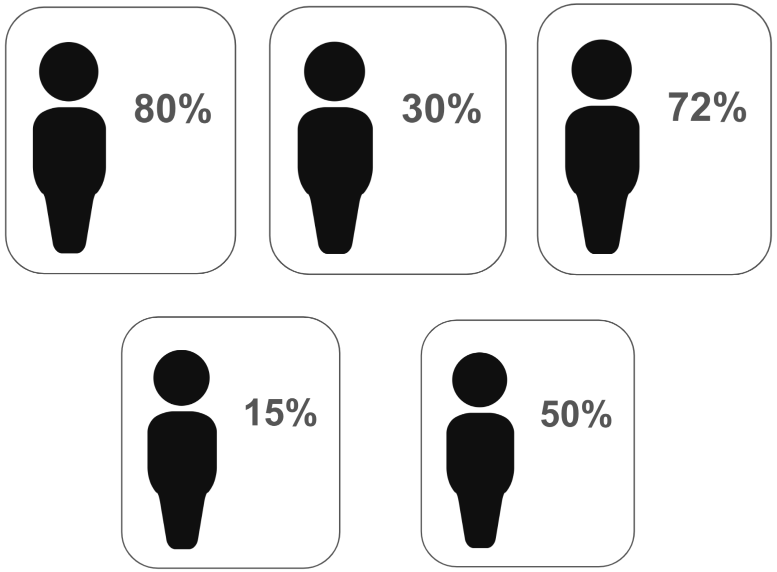
Upper Bound: (FV'98, DDFGK**O**'25)

There exists an algorithm that guarantees calibration at a rate of $O\left(T^{2/3-\epsilon}\right)$

Is calibration even necessary for predictions to be trustworthy for decisions?

Trustworthy Decision Making

Participants at COLT Workshop



Every agent wants to obtain low regret with respect to their loss function $\ell \in L$

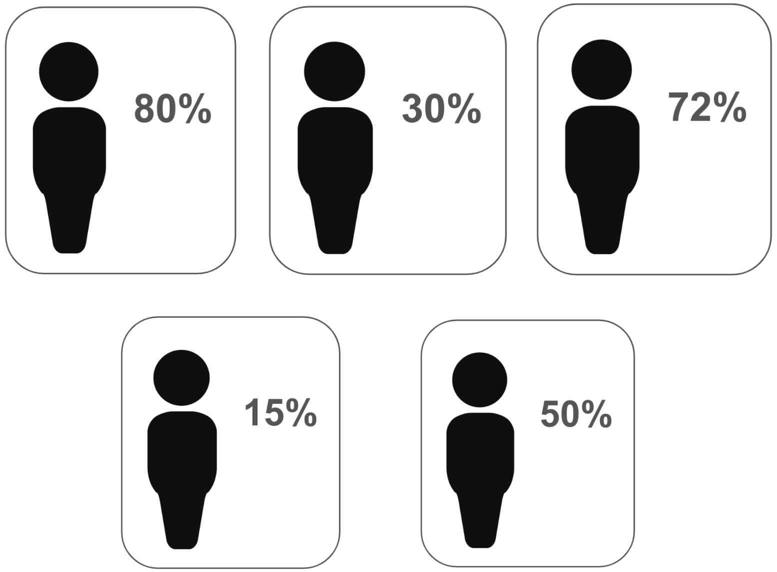
$$\sup_{\ell \in L} \sum_{t=1}^{T} \ell\left(k_\ell(p_t), y_t\right) - \min_{a \in A} \sum_{t=1}^{T} \ell\left(a, y\right) \leq o(T)$$

Theorem: (U-Calibration - KLST'23)

There exists an algorithm that makes predictions such that regret of every Agent is bounded by $O\left(\sqrt{T}\right)$

Trustworthy Decision Making

Participants at COLT Workshop



Every agent wants to obtain low regret with respect to their loss function $\ell \in L$

$$\sup_{\ell \in L} \sum_{t=1}^{T} \ell\left(k_{\ell}(p_t), y_t\right) - \min_{a \in A} \sum_{t=1}^{T} \ell\left(a, y\right) \le o(T)$$

Theorem: (U-Calibration - KLST'23)

There exists an algorithm that makes predictions such that regret of every Agent is bounded by $O\left(\sqrt{T}\right)$

Calibration is NOT necessary for predictions to be trustworthy for decisions

Decision Making with Expert Advice

Class of Experts (or Hypothesis class)

Take action A

Take action C

Take action B

In online learning, we have a class of experts to help inform our decisions.

In supervised learning, we use a hypothesis class or a class of models for decision making.

Omnipredictors

Single predictor, ***simultaneous loss minimizer*** for many losses

(Gopalan,Kalai,Reingold,Sharan,Wieder; ITCS'22)

Omniprediction: For loss class L, hypothesis class H, $\varepsilon > 0$, find predictor p such that ***for every*** $\ell \in L$

$$\mathbb{E}\left[\ell\left(k_\ell \circ p(x), y\right)\right] \le \min_{h \in H} \mathbb{E}\left[\ell\left(h(x), y\right)\right] + \varepsilon$$

k_ℓ is the "best response" function

Omnipredictors

Single predictor, ***simultaneous loss minimizer*** for many losses

(Gopalan,Kalai,Reingold,Sharan,Wieder; ITCS'22)

Omniprediction: For loss class L, hypothesis class H, $\varepsilon > 0$, find predictor p such that ***for every*** $\ell \in L$

$$\mathbb{E}\left[\ell\left(k_\ell \circ p(x), y\right)\right] \leq \min_{h \in H} \mathbb{E}\left[\ell\left(h(x), y\right)\right] + \varepsilon$$

Recovers Loss Minimization

k_ℓ is the "best response" function

Omnipredictors

Single predictor, ***simultaneous loss minimizer*** for many losses

(Gopalan,Kalai,Reingold,Sharan,Wieder; ITCS'22)

Omniprediction: For loss class L, hypothesis class H, $\varepsilon > 0$, find predictor p such that ***for every*** $\ell \in L$

$$\mathbb{E}\left[\ell\left(k_\ell \circ p(x), y\right)\right] \leq \min_{h \in H} \mathbb{E}\left[\ell\left(h(x), y\right)\right] + \varepsilon$$

No Retraining Necessary!

k_ℓ is the "best response" function

Towards Optimal Omniprediction

Key Question:

What is the Complexity of Omniprediction,
and how does it compare to Loss Minimization?

Towards Optimal Omniprediction

Key Question:

What is the Sample Complexity of Omniprediction, and how does it compare to Loss Minimization?

Optimal Loss Minimization:

$\Theta\left(d_{\ell\circ H}/\varepsilon^2\right)$

(GHKRW'23)

Omniprediction, so far:

$O\left(d_{L\circ H}/\varepsilon^6 + 1/\varepsilon^{10}\right)$

Is the gap in complexity inherent?

Towards Optimal Omniprediction

Key Question:

What is the Sample Complexity of Omniprediction, and how does it compare to Loss Minimization?

(**O.**, Kleinberg, Kim'25)

Optimal Loss Minimization:

$$\Theta\left(d_{\ell\circ H}/\varepsilon^2\right)$$

Omniprediction:

$$\Theta\left(d_{L\circ H}/\varepsilon^2\right)$$

Sample Complexity of Omniprediction ≈ Minimization of Single Loss!

Towards Optimal Omniprediction

Key Question:

What is the Sample Complexity of Omniprediction, and how does it compare to Loss Minimization?

(**O.**, Kleinberg, Kim'25)

Optimal Loss Minimization:

$\Theta\left(d_{\ell\circ H}/\varepsilon^2\right)$

Omniprediction:

$\Theta\left(d_{L\circ H}/\varepsilon^2\right)$

Similar result for RKHS by DHIPT'25

Sample Complexity of Omniprediction ≈ Minimization of Single Loss!

DHIPT'25 - From Fairness to Infinity: Outcome-Indistinguishable (Omni)Prediction in Evolving Graphs

Learning Omnipredictors

Theorem: (Gopalan,Hu,Kim,Reingold,Wieder;ITCS'23)

$$\text{Calibration} + L \circ H\text{-Multiaccuracy} \Longrightarrow \text{Omniprediction}$$

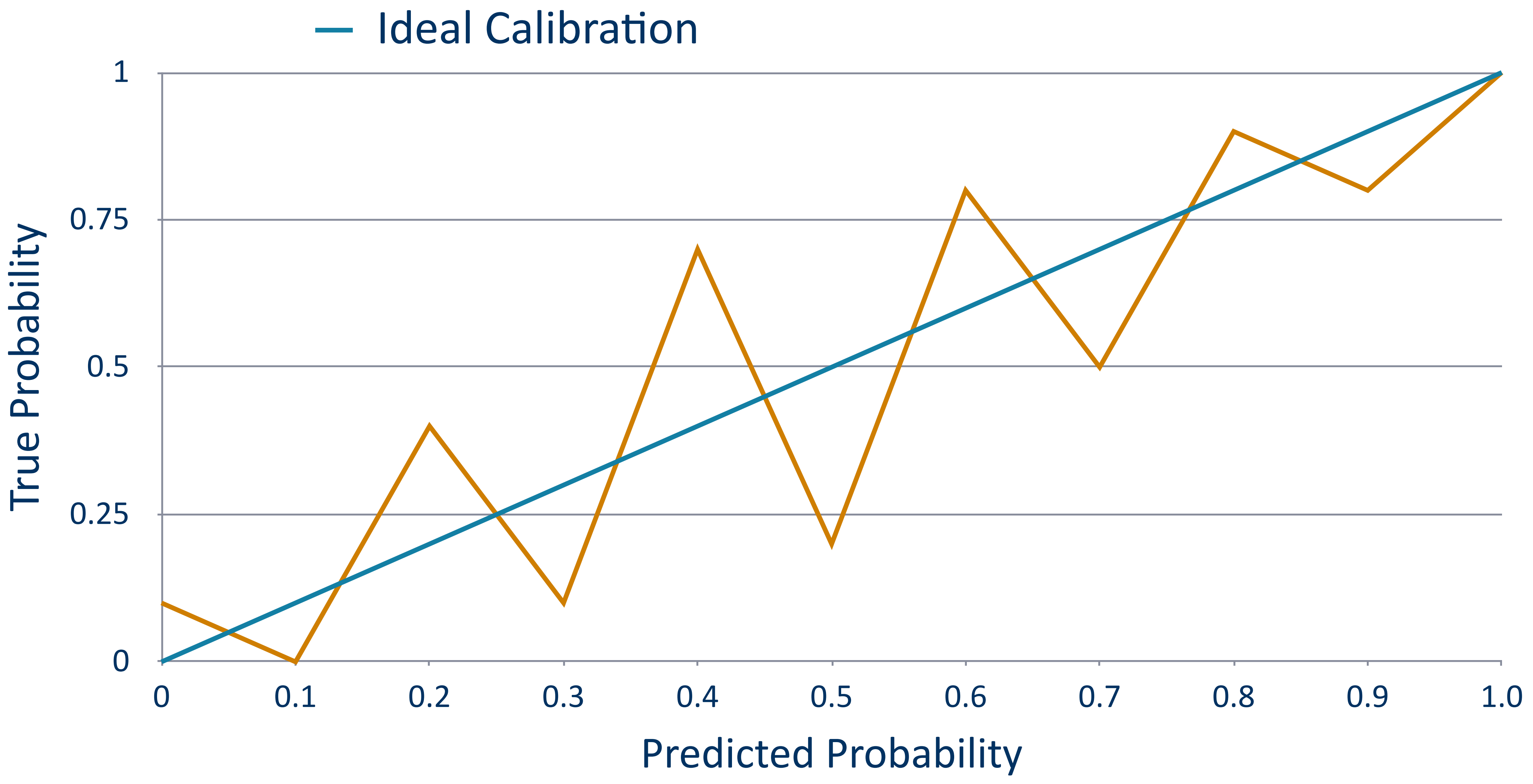
Learning Omnipredictors

Theorem: (Gopalan,Hu,Kim,Reingold,Wieder;ITCS'23)

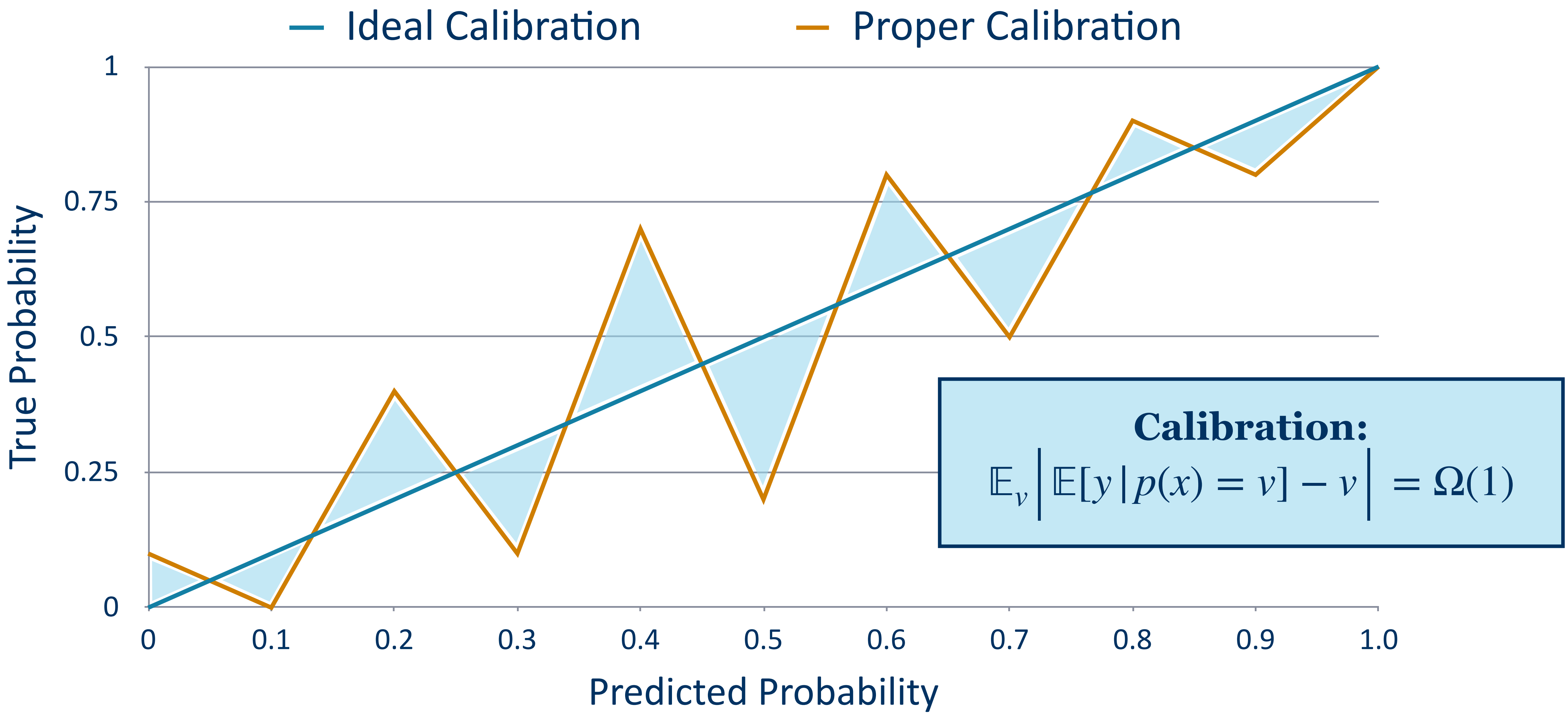
Calibration + $L \circ H$-Multiaccuracy $\implies$ Omniprediction

Theorem: (**O.**, Kleinberg, Kim'25)

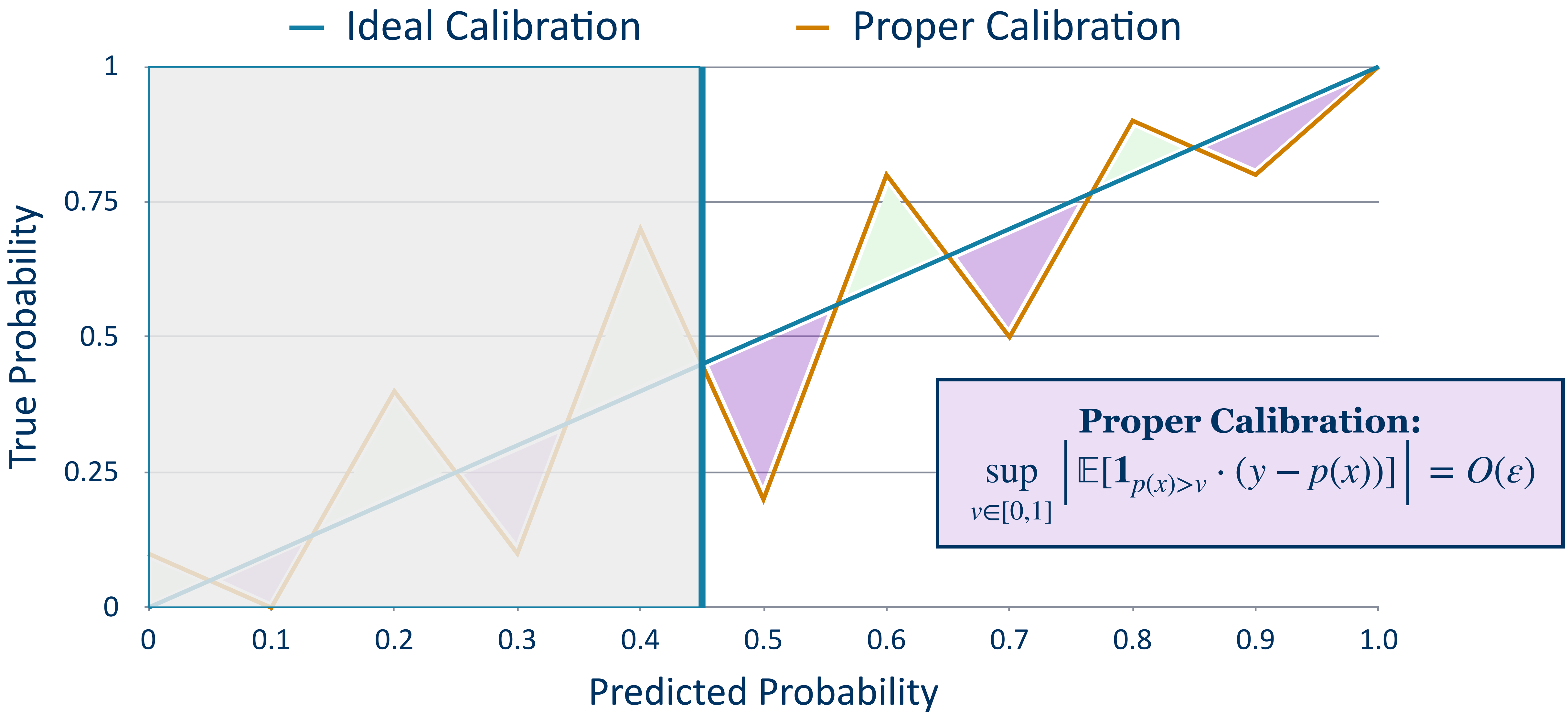
Proper Calibration + $L \circ H$-Multiaccuracy $\implies$ Omniprediction



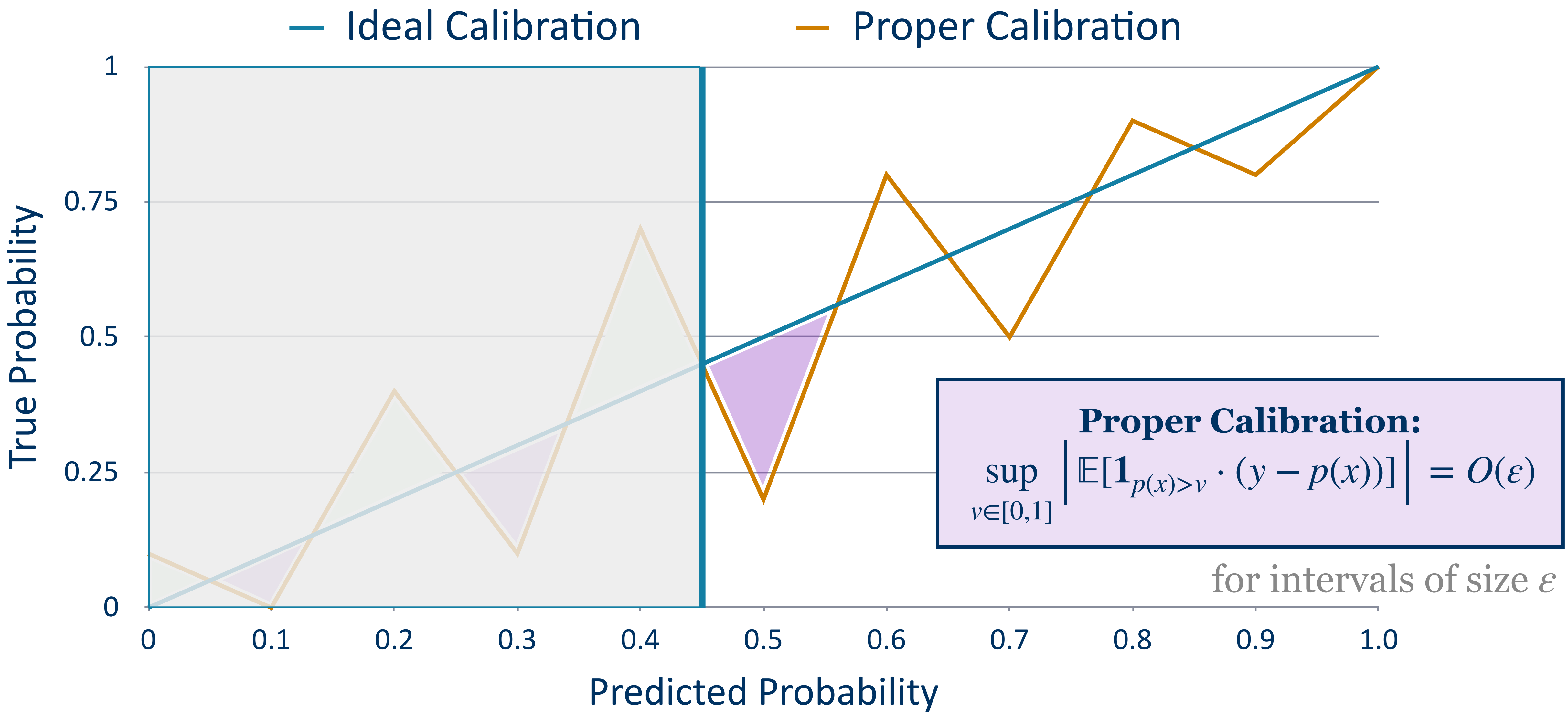
Ideal Calibration
True Probability
1
0.75
0.5
0.25
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Predicted Probability



Ideal Calibration
Proper Calibration
1
0.75
0.5
0.25
0
True Probability
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Predicted Probability
Calibration:
$\mathbb{E}_v \left| \mathbb{E}[y \mid p(x) = v] - v \right| = \Omega(1)$



Ideal Calibration
Proper Calibration
1
0.75
0.5
0.25
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
True Probability
Predicted Probability
Proper Calibration:
$$\sup_{v\in[0,1]} \left| \mathbb{E}[\mathbf{1}_{p(x)>v} \cdot (y - p(x))] \right| = O(\varepsilon)$$



Ideal Calibration
Proper Calibration
1
0.75
0.5
0.25
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
True Probability
Predicted Probability
Proper Calibration:
$\sup_{v \in [0,1]} \left| \mathbb{E}[\mathbf{1}_{p(x)>v} \cdot (y - p(x))] \right| = O(\varepsilon)$
for intervals of size ε

Achieving Proper Calibration

Theorem: (**O.**, Kleinberg, Kim'25)

Proper Calibration is efficiently testable using $O\left(1/\epsilon^2\right)$ samples

Achieving Proper Calibration

Calibration is not even statistically testable for arbitrary predictors

Theorem: (**O.**, Kleinberg, Kim'25)

Proper Calibration is efficiently testable using $O\left(1/\epsilon^2\right)$ samples

Achieving Proper Calibration

Calibration is not even statistically testable for arbitrary predictors

Theorem: (**O.**, Kleinberg, Kim'25)

Proper Calibration is efficiently testable using $O\left(1/\epsilon^2\right)$ samples

Theorem: (**O.**, Kleinberg, Kim'25)

There exists an algorithm that achieves proper calibration at a rate of $O\left(\sqrt{T}\right)$

Achieving Proper Calibration

Calibration is not even statistically testable for arbitrary predictors

Theorem: (**O.**, Kleinberg, Kim'25)

Proper Calibration is efficiently testable using $O\left(1/\epsilon^2\right)$ samples

Recall Calibration requires $\Omega(T^{0.534})$

Theorem: (**O.**, Kleinberg, Kim'25)

There exists an algorithm that achieves proper calibration at a rate of $O\left(\sqrt{T}\right)$

Achieving Proper Calibration

Calibration is not even statistically testable for arbitrary predictors

Theorem: (**O.**, Kleinberg, Kim'25)

Proper Calibration is efficiently testable using $O\left(1/\epsilon^2\right)$ samples

Recall Calibration requires $\Omega(T^{0.534})$

Theorem: (**O.**, Kleinberg, Kim'25)

There exists an algorithm that achieves proper calibration at a rate of $O\left(\sqrt{T}\right)$

Contemporary Work by (QZ'25,RSBRW'25)

QZ'25 - Truthfulness of Decision-Theoretic Calibration Measures
RSBRW'25 - Can a calibration metric be both testable and actionable?

Online Omniprediction

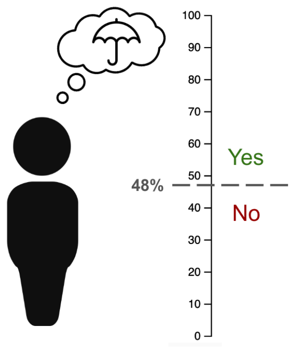
(**O.**,Kleinberg,Kim'25)

Theorem: Given a loss class L, a hypothesis class H and a sequence of adversarially chosen pairs (x_t, y_t), there exists an online algorithm that outputs predictors that achieve expected regret

$$\tilde{O}\left(\sqrt{T \cdot d^{\text{seq}}_{L \circ H}}\right)$$

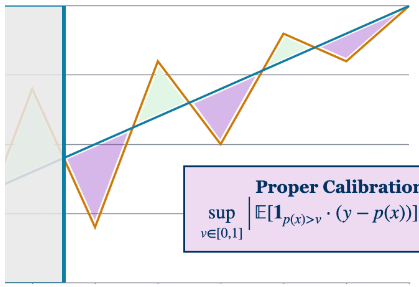
where $d_{L \circ H}$ = sequential dimension of function class derived from $L \circ H$

Key Takeaway



While Calibration provides lots of guarantees, understanding the tasks that we need it for helps us design better calibration measures for those tasks

Proper calibration is a calibration measure powerful enough to guarantee omniprediction while still efficiently achievable



Proper Calibration:

$$\sup_{v \in [0,1]} \left| \mathbb{E}[\mathbf{1}_{p(x)>v} \cdot (y - p(x))] \right| = O(\varepsilon)$$

Future Work

Some progress by LRS'25

What about settings with multiple outcomes?

Can we generalize these results to multi-class and real valued settings? Can we achieve proper calibration efficiently in higher dimensions?

For what tasks is full calibration necessary?

While calibration is sufficient for trustworthy decision making, it's not necessary.

LRS'25 - Sample Efficient Omniprediction and Downstream Swap Regret for Non-Linear Losses