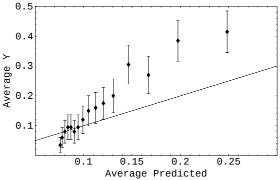
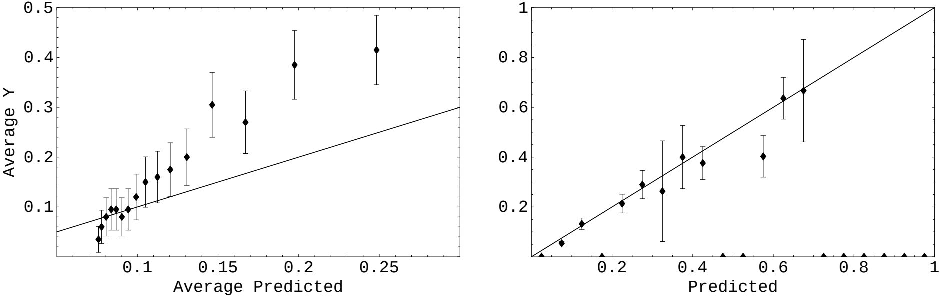


Macau, Calibeating and Fairness

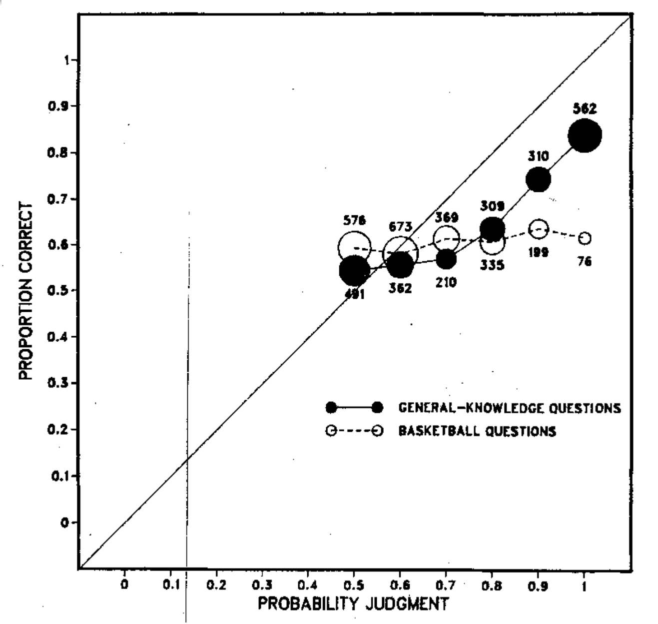
Dean P. Foster

Statistics: Anything easily fixed isn't calibrated

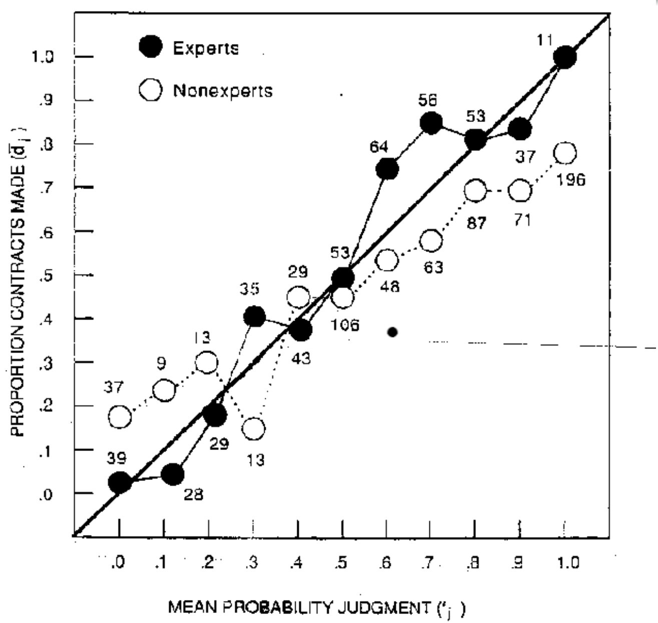


Fix the obvious problems!

Game theory: Without incentives



Game theory: With incentives!



On-line Calibration

Calibration is a minimal condition for performance

- On sequence: 0 1 0 1 0 1 0 ...
- The constant forecast of .5 is calibrated
- The constant forecast of .6 is not calibrated
- The variable forecast of .1 .9 .1 .9 .1 .9 ... is not calibrated

On-line Calibration

Calibration is a minimal condition for performance

- On sequence: 0 1 0 1 0 1 0 ...
- The constant forecast of .5 is calibrated
- The constant forecast of .6 is not calibrated
- The variable forecast of .1 .9 .1 .9 .1 .9 ... is not calibrated
    - But the forecast .1 .9 .1 .9 .1 .9 ... is pretty good!
    - Yes, it has better “refinement.”
    - But, it isn’t calibrated.

Calibration is achievable

Theorem

A calibrated forecast exists.

Calibration is achievable

Theorem

A calibrated forecast exists.

proof:

Apply mini-max theorem.

(Sergiu Hart–personal communications–1995)

Calibration is achievable

Theorem

A calibrated forecast exists.

Detailed proof:

- Game between the statistician and Nature.
- Fine the value of a $2^{2^T} \times 2^{2^T}$ matrix game.
- Happy game theorist, not so happy computational theorist.
- (Sergiu just wrote it up carefully–2023)

But that isn't what I'm going to tell you about today

But that isn't what I'm going to tell you about today

Instead: Three short talks

Which three talks

Which three talks

- First talk: Macau: Same as multi-calibration?

Which three talks

- First talk: Macau: Same as multi-calibration?
- Second talk: Calibeating: Also same as multi-calibration?

Which three talks

- First talk: Macau: Same as multi-calibration?
- Second talk: Calibeating: Also same as multi-calibration?
- Third talk: Some thoughts on fairness

Macau in one slide

- Setting: On-line decision making
  (*aka adversarial data or robust time series*)
- Goal: Use economic forecasts for decision making

Macau in one slide

- Setting: On-line decision making
  (*aka adversarial data or robust time series*)
- Goal: Use economic forecasts for decision making
- Problem: Accuracy doesn't guarantee good decisions
  (*We'll take "accuracy" = "low regret." Regret compares actual decisions to "20/20 hindsight." 100s of papers say how to get low regret.*)

Macau in one slide

- Setting: On-line decision making
  (*aka adversarial data or robust time series*)
- Goal: Use economic forecasts for decision making
- Problem: Accuracy doesn't guarantee good decisions
  (*We'll take "accuracy" = "low regret." Regret compares actual decisions to "20/20 hindsight." 100s of papers say how to get low regret.*)
- Solution: Falsifiable is better definition of error
  - you falsify a forecast by betting against it
  - The amount it loses is its *macau*.

Macau in one slide

- Setting: On-line decision making
  (*aka adversarial data or robust time series*)
- Goal: Use economic forecasts for decision making
- Problem: Accuracy doesn't guarantee good decisions
  (*We'll take "accuracy" = "low regret." Regret compares actual decisions to "20/20 hindsight." 100s of papers say how to get low regret.*)
- Solution: Falsifiable is better definition of error
  - you falsify a forecast by betting against it
  - The amount it loses is its *macau*.

Take Aways

crazy-Calibration + low-regret $\Longrightarrow$ *low-macau* $\Longrightarrow$ *good decisions*

Operationalizing falsifiability

- We will falsify someone's claim by winning bets placed against them
- Claim: $\hat{Y} \approx EY$
  - Prove it wrong by winning lots of money:

$$\text{expected winnings} = E\left(B\left(Y - \hat{Y}\right)\right)$$

  - $(Y - \hat{Y})$ is a "fair" bet
  - B is amount bet

Operationalizing falsifiability

- We will falsify someone's claim by winning bets placed against them
- Claim: $\hat{Y} \approx EY$
  - Prove it wrong by winning lots of money:

$$\text{expected winnings} = E\left(B\left(Y - \hat{Y}\right)\right)$$

  - $(Y - \hat{Y})$ is a "fair" bet
  - B is amount bet
- How to avoid being proven wrong by:

$$E\left(B\left(Y - \hat{Y}\right)\right)$$

(*Start with bet B*)

Operationalizing falsifiability

- We will falsify someone's claim by winning bets placed against them
- Claim: $\hat{Y} \approx EY$
  - Prove it wrong by winning lots of money:

$$\text{expected winnings} = E\left(B\,(Y - \hat{Y})\right)$$

  - $(Y - \hat{Y})$ is a "fair" bet
  - B is amount bet
- How to avoid being proven wrong by:

$$\text{Macau} \equiv \max_{|B|\leq 1} E\left(B\,(Y - \hat{Y})\right)$$

(*worry about worst bet*)

Operationalizing falsifiability

- We will falsify someone's claim by winning bets placed against them
- Claim: $\hat{Y} \approx EY$
  - Prove it wrong by winning lots of money:

$$\text{expected winnings} = E\left(B\left(Y - \hat{Y}\right)\right)$$

  - $(Y - \hat{Y})$ is a "fair" bet
  - B is amount bet
- How to avoid being proven wrong by:

$$\min_{\hat{Y}} \max_{|B| \leq 1} E\left(B\left(Y - \hat{Y}\right)\right)$$

(*mini-max*)

Crazy calibration variable

Y	X_1	X_2	X_3	X_4
Y_1	X_{11}	X_{12}	X_{13}	X_{14}
Y_2	X_{21}	X_{22}	X_{23}	X_{24}
Y_3	X_{31}	X_{32}	X_{33}	X_{34}
Y_4	X_{41}	X_{42}	X_{43}	X_{44}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	X_{t3}	X_{t4}

Starting with our data that we observed up to time t

Crazy calibration variable

Y	X_1	X_2	X_3	X_4
Y_1	X_{11}	X_{12}	X_{13}	X_{14}
Y_2	X_{21}	X_{22}	X_{23}	X_{24}
Y_3	X_{31}	X_{32}	X_{33}	X_{34}
Y_4	X_{41}	X_{42}	X_{43}	X_{44}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	X_{t3}	X_{t4}

$$\hat{\beta}_t = \arg\min_\beta \sum_{i=1}^{t} (Y_i - \beta' X_i)^2$$

We can fit $\hat{\beta}_t$ on everything up to time t

Crazy calibration variable

Y	X_1	X_2	X_3	X_4		
Y_1	X_{11}	X_{12}	X_{13}	X_{14}		
Y_2	X_{21}	X_{22}	X_{23}	X_{24}		
Y_3	X_{31}	X_{32}	X_{33}	X_{34}		
Y_4	X_{41}	X_{42}	X_{43}	X_{44}		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		
Y_t	X_{t1}	X_{t2}	X_{t3}	X_{t4}		
	$X_{t+1,1}$	$X_{t+1,2}$	$X_{t+1,3}$	$X_{t+1,4}$	$\hat{\beta}_t$	$\hat{Y}_{t+1} = \hat{\beta}_t' X_{t+1}$

From a new X_{t+1} we can compute $\hat{Y}_{t+1}$

Crazy calibration variable

Y	X_1	X_2	X_3	X_4	$\hat{\beta}$
Y_1	X_{11}	X_{12}	X_{13}	X_{14}	0
Y_2	X_{21}	X_{22}	X_{23}	X_{24}	$\hat{\beta}_1$
Y_3	X_{31}	X_{32}	X_{33}	X_{34}	$\hat{\beta}_2$
Y_4	X_{41}	X_{42}	X_{43}	X_{44}	$\hat{\beta}_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	X_{t3}	X_{t4}	$\hat{\beta}_{t-1}$

Looking at only the first part of the data, we can generate:

$$\hat{\beta}_0, \quad \hat{\beta}_1, \quad \hat{\beta}_2, \quad \hat{\beta}_3, \quad \hat{\beta}_4, \quad \ldots, \quad \hat{\beta}_{t-1}$$

Crazy calibration variable

Y	X_1	X_2	X_3	X_4	$\hat{\beta}$	$\hat{Y}$
Y_1	X_{11}	X_{12}	X_{13}	X_{14}	0	$\hat{Y}_1 = 0$
Y_2	X_{21}	X_{22}	X_{23}	X_{24}	$\hat{\beta}_1$	$\hat{Y}_2 = \hat{\beta}_1' X_2$
Y_3	X_{31}	X_{32}	X_{33}	X_{34}	$\hat{\beta}_2$	$\hat{Y}_3 = \hat{\beta}_2' X_3$
Y_4	X_{41}	X_{42}	X_{43}	X_{44}	$\hat{\beta}_3$	$\hat{Y}_4 = \hat{\beta}_3' X_4$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	X_{t3}	X_{t4}	$\hat{\beta}_{t-1}$	$\hat{Y}_t = \hat{\beta}_{t-1}' X_t$

Each of these leads to a next round

$$\hat{Y}_1, \quad \hat{Y}_2, \quad \hat{Y}_3, \quad \hat{Y}_4, \quad \ldots, \quad \hat{Y}_t$$

Crazy calibration variable

Y	X_1	X_2	X_3	X_4	$\hat{\beta}$	$\hat{Y}$
Y_1	X_{11}	X_{12}	X_{13}	X_{14}	0	$\hat{Y}_1 = 0$
Y_2	X_{21}	X_{22}	X_{23}	X_{24}	$\hat{\beta}_1$	$\hat{Y}_2 = \hat{\beta}_1' X_2$
Y_3	X_{31}	X_{32}	X_{33}	X_{34}	$\hat{\beta}_2$	$\hat{Y}_3 = \hat{\beta}_2' X_3$
Y_4	X_{41}	X_{42}	X_{43}	X_{44}	$\hat{\beta}_3$	$\hat{Y}_4 = \hat{\beta}_3' X_4$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	X_{t3}	X_{t4}	$\hat{\beta}_{t-1}$	$\hat{Y}_t = \hat{\beta}_{t-1}' X_t$

Theorem (F. 1991, Forster 1999)

Such an on-line least squares forecast generates low regret:

$$\sum_{t=1}^{T} (Y_t - \hat{Y}_t)^2 - \min_{\beta} \sum_{t=1}^{T} (Y_t - \beta' X_t)^2 \leq O(\log(T))$$

Crazy calibration variable

Y	X_1	X_2	X_3	X_4	$\hat{\beta}$	$\hat{Y}$
Y_1	X_{11}	X_{12}	X_{13}	X_{14}	0	$\hat{Y}_1 = 0$
Y_2	X_{21}	X_{22}	X_{23}	X_{24}	$\hat{\beta}_1$	$\hat{Y}_2 = \hat{\beta}_1' X_2$
Y_3	X_{31}	X_{32}	X_{33}	X_{34}	$\hat{\beta}_2$	$\hat{Y}_3 = \hat{\beta}_2' X_3$
Y_4	X_{41}	X_{42}	X_{43}	X_{44}	$\hat{\beta}_3$	$\hat{Y}_4 = \hat{\beta}_3' X_4$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	X_{t3}	X_{t4}	$\hat{\beta}_{t-1}$	$\hat{Y}_t = \hat{\beta}_{t-1}' X_t$

Works no matter what the X's are.

Example: Use previous $X_{t,i} = \hat{Y}_{t-i}$. (F. and Stine 2021)

But we are going to go one better: $X_t = \hat{Y}_t$.

Y	X_1	X_2	X_3	X_4	$\hat{\beta}$	$\hat{Y}$
Y_1	X_{11}	X_{12}	$\hat{Y}_1$	X_{14}	0	$\hat{Y}_1 = 0$
Y_2	X_{21}	X_{22}	$\hat{Y}_2$	X_{24}	$\hat{\beta}_1$	$\hat{Y}_2 = \hat{\beta}_1' X_2$
Y_3	X_{31}	X_{32}	$\hat{Y}_3$	X_{34}	$\hat{\beta}_2$	$\hat{Y}_3 = \hat{\beta}_2' X_3$
Y_4	X_{41}	X_{42}	$\hat{Y}_4$	X_{44}	$\hat{\beta}_3$	$\hat{Y}_4 = \hat{\beta}_3' X_4$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	$\hat{Y}_t$	X_{t4}	$\hat{\beta}_{t-1}$	$\hat{Y}_t = \hat{\beta}_{t-1}' X_t$

Theorem holds when one of the X_t's is $\hat{Y}_t$!

Crazy calibration variable

Y	X_1	X_2	X_3	X_4	$\hat{\beta}$	$\hat{Y}$
Y_1	X_{11}	X_{12}	$\hat{Y}_1$	X_{14}	0	$\hat{Y}_1 = 0$
Y_2	X_{21}	X_{22}	$\hat{Y}_2$	X_{24}	$\hat{\beta}_1$	$\hat{Y}_2 = \hat{\beta}_1' X_2$
Y_3	X_{31}	X_{32}	$\hat{Y}_3$	X_{34}	$\hat{\beta}_2$	$\hat{Y}_3 = \hat{\beta}_2' X_3$
Y_4	X_{41}	X_{42}	$\hat{Y}_4$	X_{44}	$\hat{\beta}_3$	$\hat{Y}_4 = \hat{\beta}_3' X_4$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_t	X_{t1}	X_{t2}	$\hat{Y}_t$	X_{t4}	$\hat{\beta}_{t-1}$	$\hat{Y}_t = \hat{\beta}_{t-1}' X_t$

Theorem ($\Longrightarrow$ F. and Kakade 2008, F. and Hart 2018)

Adding the crazy calibration variable generates low macau:

$$(\forall i) \quad \sum_{t=1}^{T} X_{t,i}(Y_t - \hat{Y}_t) = O(\sqrt{T \log(T)})$$

Macau as the “normal equation”

$E(Y\|X)$	Least squares	Normal equations
Statistics	$\min_{\beta} \sum (Y_i - \beta \cdot X_i)^2$	$\sum X_i \ (Y_i - \beta \cdot X_i) = 0$

The normal equation is the same as:

$$\max_{\alpha} \sum_i \alpha' X_i (Y_i - \beta' X_i)) = 0$$

Which is solved by the β minimizer:

$$\min_{\beta} \max_{\alpha} \sum_i \alpha' X_i (Y_i - \beta' X_i)) = 0$$

Macau as the "normal equation"

$E(Y\|X)$	Least squares	Normal equations
Statistics	$\min_{\beta} \sum (Y_i - \beta \cdot X_i)^2$	$\min_{\beta} \max_{\alpha} \sum \alpha \cdot X_i \ (Y_i - \beta \cdot X_i)$

Macau as the "normal equation"

$E(Y\|X)$	Least squares	Normal equations
Statistics	$\min_{\beta} \sum (Y_i - \beta \cdot X_i)^2$	$\min_{\beta} \max_{\alpha} \sum \alpha \cdot X_i \; (Y_i - \beta \cdot X_i)$
Probability	$\min_{f} E((Y - \underbrace{f(X)}_{aka\ E(Y\|X)})^2)$	$(\forall g)\ E(g(X)\ (Y - f(X))) = 0$

The normal equation is the same as:

$$\max_{g} E\,(g(X)(Y - f(X))) = 0$$

Which is solved by the $f(\cdot)$ minimizer:

$$\min_{f} \max_{g} E\,(g(X)(Y - f(X))) = 0$$

Macau as the "normal equation"

$E(Y\|X)$	Least squares	Normal equations
Statistics	$\min_\beta \sum (Y_i - \beta \cdot X_i)^2$	$\min_\beta \max_\alpha \sum \alpha \cdot X_i \ (Y_i - \beta \cdot X_i)$
Probability	$\min_f E((Y - \underbrace{f(X)}_{\text{aka } E(Y\|X)})^2)$	$\min_f \max_g E\Big(g(X) \ (Y - f(X))\Big)$

Macau as the "normal equation"

$E(Y\|X)$	Least squares	Normal equations
Statistics	$\min_\beta \sum (Y_i - \beta \cdot X_i)^2$	$\min_\beta \max_\alpha \sum \alpha \cdot X_i \; (Y_i - \beta \cdot X_i)$
Probability	$\min_f E((Y - \underbrace{f(X)}_{aka\ E(Y\|X)})^2)$	$\min_f \max_g E\Big(g(X) \; (Y - f(X))\Big)$
online	low regret	low macau

$$Regret \equiv \sum_{t=1}^{T} (Y_t - \hat{Y}_t)^2 - \min_\beta \sum_{t=1}^{T} (Y_t - \beta \cdot X_t)^2$$

Macau as the "normal equation"

$E(Y\|X)$	Least squares	Normal equations
Statistics	$\min_{\beta} \sum (Y_i - \beta \cdot X_i)^2$	$\min_{\beta} \max_{\alpha} \sum \alpha \cdot X_i \ (Y_i - \beta \cdot X_i)$
Probability	$\min_{f} E((Y - \underbrace{f(X)}_{aka\ E(Y\|X)})^2)$	$\min_{f} \max_{g} E\Big(g(X)\ (Y - f(X))\Big)$
online	low regret	low macau

$$Macau \equiv \max_{\alpha:|\alpha|\leq 1} \sum_{t=1}^{T} \alpha \cdot X_t \left(Y_t - \hat{Y}_t\right)$$

Macau as the "normal equation"

$E(Y\|X)$	Least squares	Normal equations
Statistics	$\min_\beta \sum (Y_i - \beta \cdot X_i)^2$	$\min_\beta \max_\alpha \sum \alpha \cdot X_i \; (Y_i - \beta \cdot X_i)$
Probability	$\min_f E((Y - \underbrace{f(X)}_{\text{aka } E(Y\|X)})^2)$	$\min_f \max_g E\Big(g(X) \; (Y - f(X))\Big)$
online	low regret	low macau

- statistics: Least squares $\iff$ normal equations
- probability: Least squares $\iff$ normal equations

Macau as the "normal equation"

$E(Y\|X)$	Least squares	Normal equations
Statistics	$\min_{\beta} \sum (Y_i - \beta \cdot X_i)^2$	$\min_{\beta} \max_{\alpha} \sum \alpha \cdot X_i \ (Y_i - \beta \cdot X_i)$
Probability	$\min_{f} E((Y - \underbrace{f(X)}_{\text{aka } E(Y\|X)})^2)$	$\min_{f} \max_{g} E\Big(g(X) \ (Y - f(X))\Big)$
online	low regret	low macau

Take Aways

on-line low regret $\not\Leftrightarrow$ *on-line low macau*

low regret $\not\Leftrightarrow$ low macau

No regret $\not\Rightarrow$ not falsified

t	1	2	3	4	$\cdots$	T-1	T	T+1	T+2	T+3	$\cdots$	3T
Y_t	0	0	0	0	$\cdots$	0	1	1	1	1	$\cdots$	1
X_t	1	1	1	1	$\cdots$	1	1	1	1	1	$\cdots$	1
$\hat{Y}_t$	0	0	0	0	$\cdots$	0	0	$\frac{1}{T}$	$\frac{2}{T+1}$	$\frac{3}{T+2}$	$\cdots$	$\frac{2}{3}$

How about a bet?

Not falsified $\not\Rightarrow$ no regret

t	1	2	3	4	$\cdots$	T	T+1	$\cdots$
Y_t	0	1	0	1	$\cdots$	0	1	$\cdots$
X_t	1	1	1	1	$\cdots$	1	1	$\cdots$
$\hat{Y}_t$	.6	.4	.6	.4	$\cdots$	.6	.4	$\cdots$

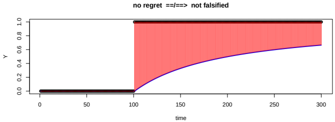
- Macau is zero
- Regret is $T/9$
- So: low macau $\not\Rightarrow$ low regret

low regret $\not\Leftrightarrow$ low macau

No regret $\not\Rightarrow$ not falsified

t	1	2	3	4	$\cdots$	T-1	T	T+1	T+2	T+3	$\cdots$	3T
Y_t	0	0	0	0	$\cdots$	0	1	1	1	1	$\cdots$	1
X_t	1	1	1	1	$\cdots$	1	1	1	1	1	$\cdots$	1
$\hat{Y}_t$	0	0	0	0	$\cdots$	0	0	$\frac{1}{T}$	$\frac{2}{T+1}$	$\frac{3}{T+2}$	$\cdots$	$\frac{2}{3}$

How about a bet?



Not falsified $\not\Rightarrow$ no regret

t	1	2	3	4	$\cdots$	T	T+1	$\cdots$
Y_t	0	1	0	1	$\cdots$	0	1	$\cdots$
X_t	1	1	1	1	$\cdots$	1	1	$\cdots$
$\hat{Y}_t$	.6	.4	.6	.4	$\cdots$	.6	.4	$\cdots$

- Macau is zero
- Regret is $T/9$
- So: low macau $\not\Rightarrow$ low regret

(Skipping these proofs)

Why is low macau useful?

$$C(a) = \sum_{t=1}^{T} c_t(a) \qquad a^* \equiv \arg\min_a C(a)$$

- Supposed each $c_t(\cdot)$ is convex
- Goal: play a to minimize $C(a)$
- Eg: We could use SGD on $\nabla c_t()$
- called "on-line convex optimization" with regret:

$$\text{regret} \;\equiv\; \sum_{t=1}^{T} (c_t(\hat{a}_t) - c_t(a^*))$$

Why is low macau useful?

$$C(a) = \sum_{t=1}^{T} c_t(a) \qquad a^* \equiv \arg\min_a C(a)$$

The regret is bounded by the gradient:

$$\begin{aligned} \text{regret} &= \sum_{t=1}^{T} (c_t(\hat{a}_t) - c_t(a^*)) \\ &\leq \sum_{t=1}^{T} (\hat{a}_t - a^*) \cdot \nabla c_t(\hat{a}_t) \end{aligned}$$

Why is low macau useful?

$$C(a) = \sum_{t=1}^{T} c_t(a) \qquad a^* \equiv \arg\min_a C(a)$$

The regret is bounded by the gradient:

$$\begin{aligned}
\text{regret} &= \sum_{t=1}^{T} (c_t(\hat{a}_t) - c_t(a^*)) \\
&\leq \sum_{t=1}^{T} (\hat{a}_t - a^*) \cdot \nabla c_t(\hat{a}_t) \\
&= \sum_{t=1}^{T} (\hat{a}_t - a^*) \cdot \left(\nabla c_t(\hat{a}_t) - \widehat{\nabla c_t}(\hat{a}_t)\right) + (\hat{a}_t - a^*) \cdot \widehat{\nabla c_t}(\hat{a}_t)
\end{aligned}$$

Why is low macau useful?

$$C(a) = \sum_{t=1}^{T} c_t(a) \qquad a^* \equiv \arg\min_a C(a)$$

The regret is bounded by the gradient:

$$\begin{aligned}
\text{regret} &= \sum_{t=1}^{T} (c_t(\hat{a}_t) - c_t(a^*)) \\
&\leq \sum_{t=1}^{T} (\hat{a}_t - a^*) \cdot \nabla c_t(\hat{a}_t) \\
&= \underbrace{\sum_{t=1}^{T} (\hat{a}_t - a^*) \cdot \left(\nabla c_t(\hat{a}_t) - \widehat{\nabla c_t}(\hat{a}_t) \right)}_{\textit{(macau!)}} + (\hat{a}_t - a^*) \cdot \underbrace{\widehat{\nabla c_t}(\hat{a}_t)}_{\textit{(zero @ } \hat{a}_t \textit{)}}
\end{aligned}$$

Why is low macau useful?

$$C(a) = \sum_{t=1}^{T} c_t(a) \qquad a^* \equiv \arg\min_a C(a)$$

The regret is bounded by the gradient:

$$\begin{aligned}
\text{regret} &= \sum_{t=1}^{T} (c_t(\hat{a}_t) - c_t(a^*)) \\
&\leq \sum_{t=1}^{T} (\hat{a}_t - a^*) \cdot \nabla c_t(\hat{a}_t) \\
&= \sum_{t=1}^{T} (\hat{a}_t - a^*) \cdot \left(\nabla c_t(\hat{a}_t) - \widehat{\nabla c_t}(\hat{a}_t) \right) + (\hat{a}_t - a^*) \cdot \widehat{\nabla c_t}(\hat{a}_t) \\
\text{regret} &\leq \text{macau}
\end{aligned}$$

Calibration Theorem

Theorem ($\Longrightarrow$ F. and Kakade 2008, $\Longleftarrow$ new)

Let R be the quadratic regret of a forecast $\hat{Y}_t$ against a linear regression on X_t. Let M be the Macau of $\hat{Y}_t$ using linear functions of X_t to create falsifying bets. Then if we have the crazy calibration variable (i.e. $[X_t]_0 = \hat{Y}_t$), then

$$R = o(T) \quad \textit{iff} \quad M = o(T).$$

Calibration Theorem

Theorem ($\Longrightarrow$ F. and Kakade 2008, $\Longleftarrow$ new)

Let R be the quadratic regret of a forecast $\hat{Y}_t$ against a linear regression on X_t. Let M be the Macau of $\hat{Y}_t$ using linear functions of X_t to create falsifying bets. Then if we have the crazy calibration variable (i.e. $[X_t]_0 = \hat{Y}_t$), then

$$R = o(T) \quad \textit{iff} \quad M = o(T).$$

Proof sketch: Consider the forecasts $(1 - w)\hat{Y}_t + w\alpha \cdot X_t$ for the *any* α. Let $Q(w)$ be the total quadratic error of this family of forecast. The following are equivalent:

- $Q(0) \leq Q(w)$ (No regret condition)
- $Q'(0)$ is zero. (No macau condition)

Calibration Theorem

Theorem ($\Longrightarrow$ F. and Kakade 2008, $\Longleftarrow$ new)

Let R be the quadratic regret of a forecast $\hat{Y}_t$ against a linear regression on X_t. Let M be the Macau of $\hat{Y}_t$ using linear functions of X_t to create falsifying bets. Then if we have the crazy calibration variable (i.e. $[X_t]_0 = \hat{Y}_t$), then

$$R = o(T) \quad \textit{iff} \quad M = o(T).$$

Note: Typically, $R = O(\log(T))$ iff $M = \tilde{O}(\sqrt{T})$ for the actual algorithms I know.

(*S. Rakhlin and D. Foster have a proof for IID.*)

Recipe for good decisions

- List bets that you would make to show $\hat{a}_t$ is not optimal
- Convert these to regression variables
- Add the crazy-calibration variable
- Run a low regret least squares algorithm
- Make decision based on this forecast

That is Macau

Take Aways

crazy-Calibration + low-regret $\iff$ *low-macau* $\implies$ *good decisions*

Second topic: Calibeating

- Predicting the "grand average" is calibrated
  - But it is a crappy forecast.
- We have lots of ways of generating good forecasts:
  - probabilistic models
  - Time series: ARIMA, etc
  - on-line least squares regression
  - Combining experts
- None are guaranteed to be calibrated

Second topic: Calibeating

- Predicting the “grand average” is calibrated
  - But it is a crappy forecast.
- We have lots of ways of generating good forecasts:
  - probabilistic models
  - Time series: ARIMA, etc
  - on-line least squares regression
  - Combining experts
- None are guaranteed to be calibrated

Goal: Find a way to convert these good forecasts into calibrated forecasts without removing their goodness.

On-line Calibration

Recall our “good” by not calibrated forecast from the introduction:

- On sequence: 0 1 0 1 0 1 0 ...
- The constant forecast of .5 is calibrated
- The variable forecast of .1 .9 .1 .9 .1 .9 ... is not calibrated
  - It has better fit: called “refinement.”
  - But, it isn’t calibrated.
  - Our goal: Keep this refinement, but make it calibrated

Bias / Variance decomposition

- bias:

$$\beta \equiv E(Y|\hat{Y}) - \hat{Y}$$

- variance:

$$\text{VAR} = \text{Var}(Y - E(Y|\hat{Y}))$$

- Mean Squared error:

$$\text{MSE} = E(Y - \hat{Y})^2 = E(\beta^2) + \text{VAR}$$

- For binary sequences:
  - Bias is called *Calibration*
  - Variance is called *Refinement*
  - MSE is called *Brier Score*

Brier score

- "Conditional expectation":

$$\rho(x) = \frac{\sum_t Y_t \, I_{\hat{y}_t = x}}{\sum I_{\hat{y}_t = x}}$$

- Bias: $\beta(x) = \rho(x) - x$
- Brier score / MSE:

$$BS = \frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{Y}_t)^2$$

- Decomposition (MSE = bias + Variance):

$$\underbrace{\frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{Y}_t)^2}_{BS} = \underbrace{\frac{1}{T}\sum^{T}(\hat{Y} - \rho(\hat{Y}))^2}_{Calibration} + \underbrace{\frac{1}{T}\sum^{T}(Y_t - \rho(\hat{Y}_t))^2}_{Refinement}$$

Defining calibeating

Calibration is fixable after the fact.

- But, can we fix it as we go along?
- Start with a forecast $\hat{y}_t$
- Calibration $K(\hat{y})$
- Refinement $R(\hat{y})$

Find a new forecast $\tilde{y}_t$ that doesn't pay the calibration costs of $\hat{y}$

Definition (Calibeating)

$\tilde{y}$ calibeats $\hat{y}$ if:

$$\mathrm{BS}(\tilde{y}) \leq R(\hat{y}).$$

- $\tilde{y}$ keeps any patterns found by $\hat{y}$
- $\tilde{y}$ doesn't "pay" the calibration error

Calibeating many forecasters

We can extend this to calibeating many forecasters.

Definition (Calibeating)

$\tilde{y}$ calibeats a collection of forecasts $\{\hat{y}^1, \ldots, \hat{y}^n\}$ if for all i:

$$\mathrm{BS}(\tilde{y}) \leq R(\hat{y}^i).$$

Calibeating is easy

- Algorithm to calibeat a family of forecasts: $\hat{y}_t^i$
  - Break up the interval $[0, 1]$ into small buckets B_j.
  - Intersect the buckets
  - Compute the average on each bucket

Theorem

The forecast combination $\tilde{y}_t$ will ϵ-calibeat $\hat{y}_t^i$ if we use buckets with width less than ϵ.

Calibeating is easy, but it can be calibeaten!

We can find $\tilde{y}$ that calibeats $\hat{y}$. But, there is no reason for $\tilde{y}$ to be calibrated. So it can be calibeaten. The result likewise isn't calibrated, so it can be calibeaten.

Calibeating is easy, but it can be calibeaten!

We can find $\tilde{y}$ that calibeats $\hat{y}$. But, there is no reason for $\tilde{y}$ to be calibrated. So it can be calibeaten. The result likewise isn't calibrated, so it can be calibeaten.

- This can go on ad infinitum

Stopping the infinite regress

We can have C_t calibeat A_t and B_t.

- Suppose at each time t we pick $B_t = C_t$.
- Requires a fixed point computation
- C_t calibeats A_t
- C_t calibeats C_t:

$$BS(C_t) \leq R(C_t)$$

So C_t is calibrated.

Theorem

For any set of forecasts, there is a combination forecast which calibeats each element in the set, and is also calibrated.

Freebie: Calibeating yourself is calibrated

If we use this theorem with an empty set then C is calibrated:

Corollary

If C calibeats itself, then C is calibrated.

About fixed points

Suppose we will forecast C_t. The calibeating algorithm would say we should instead forecast $g(A_t, C_t)$. If this happens to be C_t, we are done. Ignoring A_t this means we want $C_t = g(C_t)$.

About fixed points

Suppose we will forecast C_t. The calibeating algorithm would say we should instead forecast $g(A_t, C_t)$. If this happens to be C_t, we are done. Ignoring A_t this means we want $C_t = g(C_t)$.

Theorem (Outgoing distribution)

There exists a probability distribution on C such that:

$$E(|x - C|^2 - |x - g(C)|^2) \leq \delta^2$$

for all x.

Proof is via the mini-max theorem (so linear programming can find the answer.)

- This means the BS of using C is better than the BS of using the correct answer $g(C)$.

True fixed points

Theorem (Outgoing fixed point)

For any smooth $g()$ and any closed convex set $\mathcal{S}$, there exists a point $C \in \mathcal{S}$ such that:

$$E(|x - C|^2 - |x - g(C)|^2) \leq 0$$

for all $x \in \mathcal{S}$.

Proof is via the Brouwer's fixed point. In fact, it is equivalent to Brouwer's fixed point theorem.

True fixed points

Theorem (Outgoing fixed point)

For any smooth $g()$ and any closed convex set $\mathcal{S}$, there exists a point $C \in \mathcal{S}$ such that:

$$E(|x - C|^2 - |x - g(C)|^2) \leq 0$$

for all $x \in \mathcal{S}$.

- Can create a deterministic “weak” calibration

True fixed points

Theorem (Outgoing fixed point)

For any smooth $g()$ and any closed convex set $\mathcal{S}$, there exists a point $C \in \mathcal{S}$ such that:

$$E(|x - C|^2 - |x - g(C)|^2) \leq 0$$

for all $x \in \mathcal{S}$.

- Using rounding, it can create a local random calibrated forecast
  - Randomly round to nearest grid point
  - First few digits aren't random, just the least significant one
  - Need this minimal amount of rounding to avoid impossibility result mentioned this morning

True fixed points

Theorem (Outgoing fixed point)

For any smooth $g()$ and any closed convex set $\mathcal{S}$, there exists a point $C \in \mathcal{S}$ such that:

$$E(|x - C|^2 - |x - g(C)|^2) \leq 0$$

for all $x \in \mathcal{S}$.

- Fixed points are hard to find
- Basically need to do exhaustive search at every time period
- CS people call complexity class PPAD

Forms of calibeating

We've have four forms of calibeating:

simple	Distribution	local random	deterministic
LS or average	LP	Fixed point	Fixed point
~~calibrated~~	classic calibration	Both classic and weak	Weak
quadratic safe	Not quadratic safe	quadratic safe	quadratic safe

Final topic: Thoughts on what to calibrate

Fairness and incentives

- Consider predicts used for college admissions
  - We'll call the prediction: SAT
  - We'll call the Y variable: GPA
- We are interested in fair incentives
  - The incentive story works better for employment,
  - But the names will be useful, so we'll stick with college admissions

Regress Y on X or regression X on Y?

- Basic discrimination:

$$E(\text{GPA}|\text{blue, SAT=x}) > E(\text{GPA}|\text{orange, SAT=x})$$

  - Better off being orange
  - Richard Posner argued economics would drive it out
  - So it simply doesn't exist due to "rationality"

Regress Y on X or regression X on Y?

- Basic discrimination:

$$E(\text{GPA}|\text{blue, SAT=x}) > E(\text{GPA}|\text{orange, SAT=x})$$

  - Better off being orange
  - Richard Posner argued economics would drive it out
  - So it simply doesn't exist due to "rationality"
- But even if

$$E(\text{GPA}|\text{blue, SAT=x}) = E(\text{GPA}|\text{orange, SAT=x})$$

  we might have:

$$E(\text{SAT}|\text{blue, skill=y}) < E(\text{SAT}|\text{orange, skill=y})$$

- So still better off being Orange!

Backwards regression

- Traditional regression:

$$\min_f E\left((Y - f(X))^2\right)$$

- Reverse regression:

$$\min_g E\left((g(Y) - X)^2\right)$$

- Even if $f()$ and $g()$ are linear, $f \neq g^{-1}$
- (unless we have a perfect fit)
- Called regression to the mean

No measurement of skill

- We don't have skill, but we do have GPA
- So, regress SATs on GPAs and make that calibrated
  - Fair incentives
  - Economics won't come to this solution with Laissez-faire
  - Needs government intervention (F. and Vohra, 1992)

No measurement of skill

- We don't have skill, but we do have GPA
- So, regress SATs on GPAs and make that calibrated
  - Fair incentives
  - Economics won't come to this solution with Laissez-faire
  - Needs government intervention (F. and Vohra, 1992)
- Fairness then is best approximated by:

$$E(\text{SAT}|\text{blue, GPA=y}) \approx E(\text{SAT}|\text{orange, GPA=y})$$

References: Three different Fosters

Me:

- — (1991) "Prediction in the worst case."
- — and R. Vohra (1991-1998) "Asymptotic Calibration."
- — and R. Vohra (1992) "...Affirmative Action."
- — and S. Kakade "Deterministic calibration and Nash."
- — and S. Hart (2021) "...Leaky forecasts" (easier reading).
- — and S. Hart (2022) "Calibeating."
- — and R. Stine (2021) "Martingales and forecasts."

Dylan:

- Dylan Foster and Sasha Rakhlin (2021) "SquareCB."

Jürgen:

- J. Forster (1999) "...Linear Regression."

1:

Take Aways

crazy-Calibration + low-regret $\iff$ *low-macau*

1:

Take Aways

crazy-Calibration + low-regret $\iff$ *low-macau*

2:

simple	Distribution	local random	deterministic
LS or average	LP	Fixed point	Fixed point
~~calibrated~~	classic calibration	Both classic and weak	Weak

1:

Take Aways

crazy-Calibration + low-regret $\iff$ *low-macau*

2:

simple	Distribution	local random	deterministic
LS or average	LP	Fixed point	Fixed point
~~calibrated~~	classic calibration	Both classic and weak	Weak

3: Calibrate SATs given GPAs

1:

Take Aways

crazy-Calibration + low-regret $\Longleftrightarrow$ *low-macau*

2:

simple	Distribution	local random	deterministic
LS or average	LP	Fixed point	Fixed point
~~calibrated~~	classic calibration	Both classic and weak	Weak

3: Calibrate SATs given GPAs

Thanks!