A User Friendly Tool for Deriving Online Learning Algorithms

Aaron Roth

Penn

Based on joint work with
Daniel Lee, Georgy Noarov and Mallesh Pai

1

Two motivating online adversarial problems

2

What do/should reported uncertainty estimates mean?

Given your features x, our model predicts your expected disease severity in two days time is $f(x)$.

How sure are you?

I have a 95% prediction interval that your severity will be in $[\ell(x), u(x)]$.

Hmmm…

3

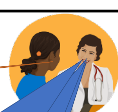
Marginal Guarantees.

$[\ell(x), u(x)]$ is a 95% marginal prediction interval.

But I'm part of a demographic group representing less than 5% of the population…

5

Marginal Guarantees.



What about for people like me?

For African Americans under the age of 50 the 95% prediction interval is [a,b]

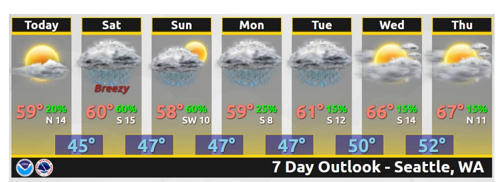
For women with a family history of diabetes the 95% prediction interval is [c,d]

What does this mean for me?

For people with egg allergies and no history of smoking, the 95% prediction interval is [e, f].

6

Calibration



7

Mean Multicalibration

Batch: [Hebert-Johnson, Kim, Reingold, Rothblum '18]
Online: [Gupta, Jung, Noarov, Pai, Roth '22]

"A sequence of predictions y_t is multicalibrated on a sequence of examples (x_t, y_t) with respect to a set of demographic groups G if for every $S \in G$ the predictions are calibrated on the subsequence:
$\{(x_t, y_t): x_t \in S\}$"

8

Prediction Interval Multivalidity

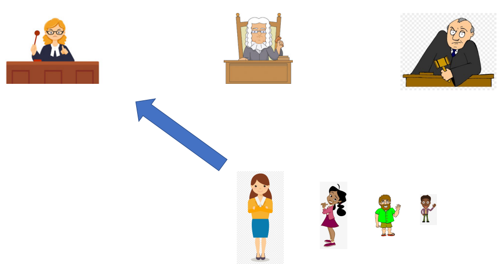
[Gupta, Jung, Noarov, Pai, Roth '22]

"A sequence of 95%-prediction intervals $[\ell_t, u_t]$ is multivalid on a sequence of examples (x_t, y_t) with respect to a set of demographic groups G if for every $S \in G$ and for every interval $[\ell, u]$, the prediction intervals cover 95% of the labels in the set:
$\{(x_t, y_t): x_t \in S, [\ell_t, u_t] \approx [\ell, u]\}$"

9

Multi-Group Optimal Assignment

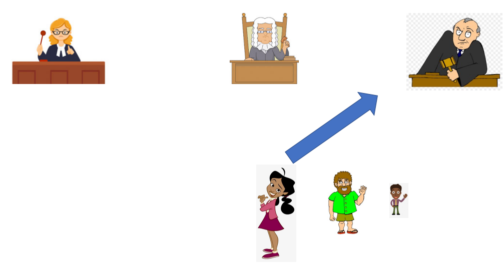
[Blum, Lykouris '20]



10

Multi-Group Optimal Assignment

[Blum, Lykouris '20]



11

Multi-Group Optimal Assignment

[Blum, Lykouris '20]

- Individuals x_t (arriving in sequence) can belong to multiple groups $g \in G$.
- As they arrive, we need to assign individuals to judges $j \in m$, who will make some impactful prediction $\hat{y}_j^t$.
- Different judges decide differently, and might have different error rates on different demographic groups:

$$err(j, g) = \sum_{t: x_t \in g} 1[y_t \neq \hat{y}_j^t].$$

- Goal: Assign people so that (up to diminishing regret terms) the average error on each group $g \in G$ is as low as it would have been had we assigned everyone in g to judge $j^* = \arg\min_j err(j, g)$

12

Other Problems

- Expert Learning Problems
  - No external regret
  - No internal regret
  - No adaptive regret
  - No regret to sleeping experts
  - No subsequence Regret
  - …
- Calibration problems
  - Mean (multi)-calibration
  - Variance (multi)-calibration
  - Prediction interval multi-validity
  - …
- Fast Blackwell Approachability
- …

13

A Simple Unifying Framework:
"Online Minimax Multiobjective Optimization"

- In rounds $t = 1, \dots, T$:
  1. The adversary proposes an *environment* consisting of:
     1. Convex compact action sets X^t and Y^t for the learner and adversary, and
     2. A vector valued loss function $\ell^t: X^t \times Y^t \to [-1,1]^d$ that in every coordinate is convex in its first argument and concave in its second.
  2. The learner selects action $x^t \in X^t$.
  3. Observing this, the adversary selects $y^t \in Y^t$ in response.
  4. The learner suffers (and observes) loss vector $\ell^t(x^t, y^t) \in [-1,1]^d$

The learner's goal is to minimize:

$$\max_{j \in [d]} \sum_{t=1}^{T} \ell_j^t(x^t, y^t)$$

14

An Interlude: Zero Sum Games

- A Zero Sum Game is defined by:
  1. A *minimization player* (the learner) with convex compact strategy space A_1
  2. A *maximization player* (the adversary) with convex compact strategy space A_2
  3. A utility function $u: A_1 \times A_2 \to \mathbb{R}$ that is convex in its first coordinate and concave in its second.

- Sion's Minimax Theorem:

$$\min_{a_1 \in A_1} \max_{a_2 \in A_2} u(a_1, a_2) = \max_{a_2 \in A_2} \min_{a_1 \in A_1} u(a_1, a_2)$$

"Order of play doesn't matter"

15

Can we just solve zero sum games?

- First idea: Just set $u(x, y) = \max_j \ell_j^t(x, y)$?
- Doesn't work --- the max does not preserve concavity for the adversary.
- The minimax theorem really doesn't hold. E.g.
  - $X^t = Y^t = \Delta[d]$, and $u(P_1, P_2) = (P_2[i] - P_1[i])_{i=1}^d$
    - Then if Max goes first, Min can obtain payoff 0.
    - But if Min goes first, Max can guarantee payoff $1 - \frac{1}{d}$.

16

What can we hope for?

- Two values for the game:
  - The Learner Moves First (LMF) Value:
    $$w_L^t = \min_{x^t} \max_{y^t} \left(\max_{j \in [d]} \ell_j^t(x^t, y^t) \right)$$
  - The Adversary Moves First (AMF) Value:
    $$w_A^t = \max_{y^t} \min_{x^t} \left(\max_{j \in [d]} \ell_j^t(x^t, y^t) \right)$$
  - We know $w_L^t > w_A^t$
- In reality, learner moves first…
- But we want diminishing regret to the AMF value:

$$\max_{j \in [d]} \left(\frac{1}{T} \sum_{t=1}^{T} \ell_j^t(x^t, y^t) \right) \le \frac{1}{T} \sum_{t=1}^{T} w_A^t + o(1)$$

17

We can achieve this.
Main Idea:

1. Define a *surrogate loss function*.
   1. Define the AMF Regret as: $R_j^t = \sum_{s=1}^{t} \left(\ell_j^s(x^s, y^s) - w_A^s \right)$, $R^t = \max_j R_j^t$
   2. The surrogate loss is $L^t = \sum_j \exp(\eta R_j^t)$
   3. The surrogate loss bounds AMF regret: $R^T \le \frac{\log(L^T)}{\eta}$
2. Observe that the surrogate telescopes:
   $$L^t \le (4\eta^2 + 1) L^{t-1} + \eta \sum_j \exp(\eta R_j^{t-1}) (\ell_j^t(x^t, y^t) - w_A^t)$$
3. Define a surrogate game:
   $$u^t(x^t, y^t) = \sum_j \exp(\eta R_j^{t-1}) (\ell_j^t(x^t, y^t) - w_A^t)$$
   - Surrogate game is convex/concave!

18

We can achieve this.
Main Idea:

$$u^t(x^t, y^t) = \sum_j \exp(\eta R_j^{t-1}) (\ell_j^t(x^t, y^t) - w_A^t)$$

1. The value of the surrogate game is 0
   - Minimax holds, so we imagine adversary goes first. By definition, learner can obtain the AMF value w_A^t
2. If the learner plays the minimax equilibrium of the surrogate game at every round, we guarantee:
   $$\max_{j \in [d]} \left(\frac{1}{T} \sum_{t=1}^{T} \ell_j^t(x^t, y^t) \right) \le \frac{1}{T} \sum_{t=1}^{T} w_A^t + 4\sqrt{\frac{\ln(d)}{T}}$$

19

Instantiating the Generic Framework
Multicalibration

- A discrete action space for the learner $X = \left\{0, \frac{1}{r}, \frac{2}{r}, \ldots, 1\right\}$
- A dimension in the loss function for each group $g \in G$, prediction $p \in X$ and *direction* $\sigma \in \{-1,1\}$:

$$\ell^t_{(g,p,\sigma)}(p^t, y^t) = 1[x^t \in g] \cdot 1[p^t = p] \cdot \sigma(p - y^t)$$

- Observe that $w^t_A \leq \frac{1}{r}$
  - Can match y^t up to discretization error $\pm\frac{1}{r}$
- Invoke the generic bound:

Simultaneously for every group $g \in G$ and prediction $p \in X$:

$$\frac{1}{T}\left|\sum_{t:x^t \in g, p^t = p} (y^t - p)\right| \leq \frac{1}{r} + 4\sqrt{\frac{\ln(2r \cdot |G|)}{T}}$$

20

Deriving the Algorithm
Compute the Minimax Equilibrium of the Surrogate Game

A simple, efficient closed form

- For t = 1 to T:
  - Compute $C^p_{t-1}(x^t) = \sum_{g \in G(x^t)} \exp\left(\eta R^{t-1}_{g,p}\right) - \exp\left(-\eta R^{t-1}_{g,p}\right)$ for p $\in [r]$
  - If $C^p_{t-1}(x_t) > 0$ for all p then predict $p^t = 1$
  - If $C^p_{t-1}(x_t) < 0$ for all p then predict $p^t = 0$
  - Otherwise:
    - find p^* s.t. $C^{p^*}_{t-1}(x_t) \cdot C^{p^*+1}_{t-1}(x_t) \leq 0$
    - Let $q_t \in [0,1]$ be s.t. $q_t \cdot C^{p^*}_{t-1}(x_t) + (1 - q_t) \cdot C^{p^*+1}_{t-1}(x_t) = 0$
    - Predict $p^t = p^*$ with probability q_t, otherwise predict $p^t = p^* + \frac{1}{r}$

21

Instantiating the Generic Framework
Multigroup Optimal Assignment

- Define a dimension in the loss function for each group $g \in G$ and judge $j' \in [m]$:

$$\ell^t_{(g,j')}(j, (\hat{y}, \hat{y}_1, \ldots \hat{y}_m)) = 1[x_t \in g] \cdot \left(1\left[\hat{y}^t_j \neq y_t\right] - 1\left[\hat{y}^t_{j'} \neq y_t\right]\right)$$

- Observe that $w^t_A = 0$
  - (Once error rates are fixed, assign x_t to $j^* = \arg\min_j \Pr[\hat{y}^t_j \neq y_t]$)
- Invoke the generic bound:

Simultaneously for every group $g \in G$:

$$\frac{1}{T}\sum_{t:x^t \in g} \Pr\left[\hat{y}^t_{j^t} \neq y_t\right] \leq \frac{1}{T}\min_{j^*(g) \in [m]} \sum_{t:x^t \in g} \Pr\left[\hat{y}^t_{j^*(g)} \neq y_t\right] + 4\sqrt{\frac{\ln(m \cdot |G|)}{T}}$$

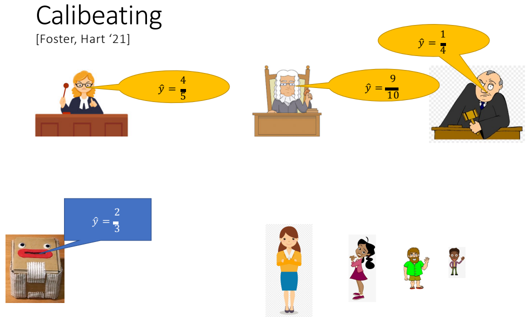
22

Can combine compatible constraints!

- Combining multigroup regret and calibration:
  - If the loss function used in regret is *proper*…
    - *E.g.* $\ell(p,y) = (p-y)^2$, $\ell(p,y) = y \cdot \ln(p) + (1-y) \cdot \ln(1-p)$
  - Then fixing an adversary's distribution over labels, the loss minimizing prediction is its expected value (so also satisfies calibration constraints)
- So regret measured with respect to a proper scoring rule is compatible with calibration
  - Immediately yields algorithms for "Multi-Calibeating"

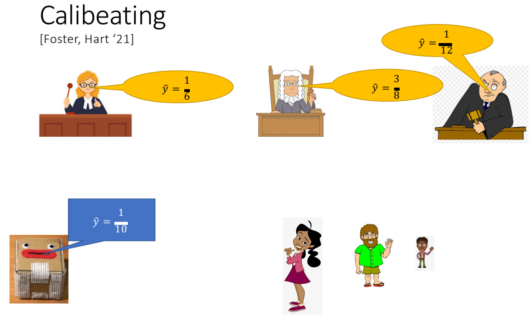
24

Calibeating
[Foster, Hart '21]



25

Calibeating
[Foster, Hart '21]



26

Calibeating

[Foster, Hart '21]

- There is an arbitrary collection of m models $f_i: X \to [0,1]$
- Each round, an arbitrary context $x^t \in X$ arrives. The models produce predictions $f_1(x^t), \ldots, f_m(x^t)$.
- The algorithm produces a prediction $p^t \in [0,1]$ and learns $y^t \in [0,1]$.
- Goal: Predictions p^t should be calibrated *and* for every i:

$$\sum_t (p^t - y^t)^2 \leq \sum_t (f_i(x^t) - y^t)^2 + o(T)$$

- *In fact, want to *strictly* improve by calibration error of f_i.

27

Combining Constraints in Our Framework: (Multi)-Calibeating

Given an arbitrary collection of groups G and an arbitrary collection of m forecasters f_i, invoking our bounds gives an algorithm that promises that for every group $g \in G$ and every model f_i,

$$\frac{1}{T}\sum_{t:x^t \in g} (p^t - y^t)^2 \leq \frac{1}{T}\sum_{t:x^t \in g} (f_i(x^t) - y^t)^2 - \mathrm{CalError}(f_i) + 4\sqrt{\frac{\ln(2m \cdot |G|)}{T}}$$

And for every prediction p:

$$\frac{1}{T}\left|\sum_{t:x^t \in g, p^t = p} (y^t - p)\right| \leq 4\sqrt{\frac{\ln(2m \cdot |G|)}{T}}$$

(Compared to FH'21, get an exponentially improved dependence on m and get subgroup guarantees)

28

Lots of other problems

Optimal bounds (via application of the main theorem), and efficient algorithms via equilibrium computation.

- No external regret
- No internal regret
- No adaptive regret
- No regret to sleeping experts
- No subsequence regret
- Mean Conditioned Moment (multi)-calibration
- Multivalid Prediction Intervals
- Fast Polytope Blackwell Approachability
- …

(any problem expressible as satisfying a finite number of linear constraints on average)

30

Thanks!

Online Minimax Multiobjective Optimization: Multicalibeating and Other Applications.

Daniel Lee, Georgy Noarov, Mallesh Pai, Aaron Roth. Manuscript, 2022

Online Multivalid Learning: Means, Moments, and Prediction Intervals.

Varun Gupta, Chris Jung, Georgy Noarov, Mallesh Pai, Aaron Roth. ITCS 2022

31