MULTICALIBRATION, UNIVERSAL ADAPTABILITY AND CAUSALITY

Omer Reingold

Stanford

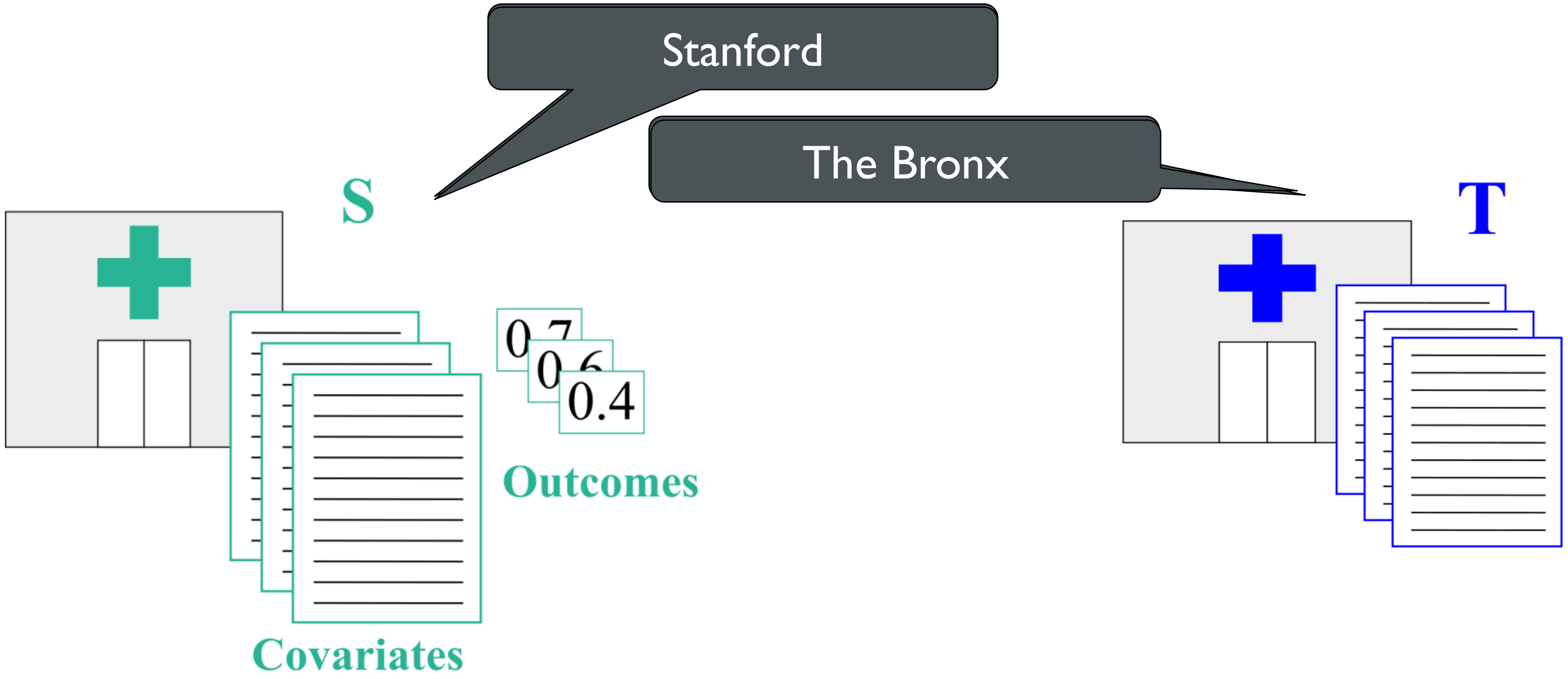
Joint with: Michael Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter

+ Work in Progress + Speculative Musings

LET'S TALK ABOUT IT

- Source to target adaptation through
  - Propensity score reweighing (prevalent paradigm).
- Universal Adaptation – from one source to many targets.
  - Multicalibration (fairness) to the rescue.
- Ho yes, it’s a semester on causality.
  - Propensity scoring for treatment effect estimations.
  - Musings on multicalibration and causality.

GENERALIZATIONS FROM A SOURCE TO A TARGET



STATISTICAL ESTIMATION SETUP

- $X \in \mathcal{X}$ — covariates (features of individuals)
- $Y \in \mathcal{Y} \subseteq [0,1]$ — outcome of interest (real or discrete)
- $Z \in \{s, t\}$ — source vs. target population (think uniform)
- Goal: target estimation $E[Y|Z = t]$ (applies to other statistics)

KEY ASSUMPTIONS

- Conditional independence: Y and Z are independent conditioned on X (the rule/correlation we are trying to learn is the same in source and destination).
- Sufficient representation: every (large) target subpopulation somewhat represented in source.
- Both are required for our work (universal adaptability) but also for propensity score reweighing.

PROPENSITY SCORE

- Models the shift in distributions of covariates.
- Odds of sampling a given individual X from source vs. target
- Propensity Score: $e_{st}(x) = Pr[Z = s|X = x]$
  - Positivity: $e_{st}(x)$ is bounded away from 0
  - Note: $1 - e_{st}(x) = Pr[Z = t|X = x]$

PROPENSITY SCORE $\Rightarrow$ VALID INFERENCE

- Under conditional independence:

$$E[Y|Z=t] = E\left[\left(\frac{1-e_{st}(X)}{e_{st}(X)}\right)\cdot Y|Z=s\right]$$

- Propensity score reweighing:
  - Estimate $e_{st}(x)$ using unlabeled samples $\{X_i\} \sim s$ and $\{X_i\} \sim t$
  - Reweigh labeled source samples $\{(X_i, Y_i)\} \sim s$ by propensity odds $(1 - e_{st}(X_i))/e_{st}(X_i)$

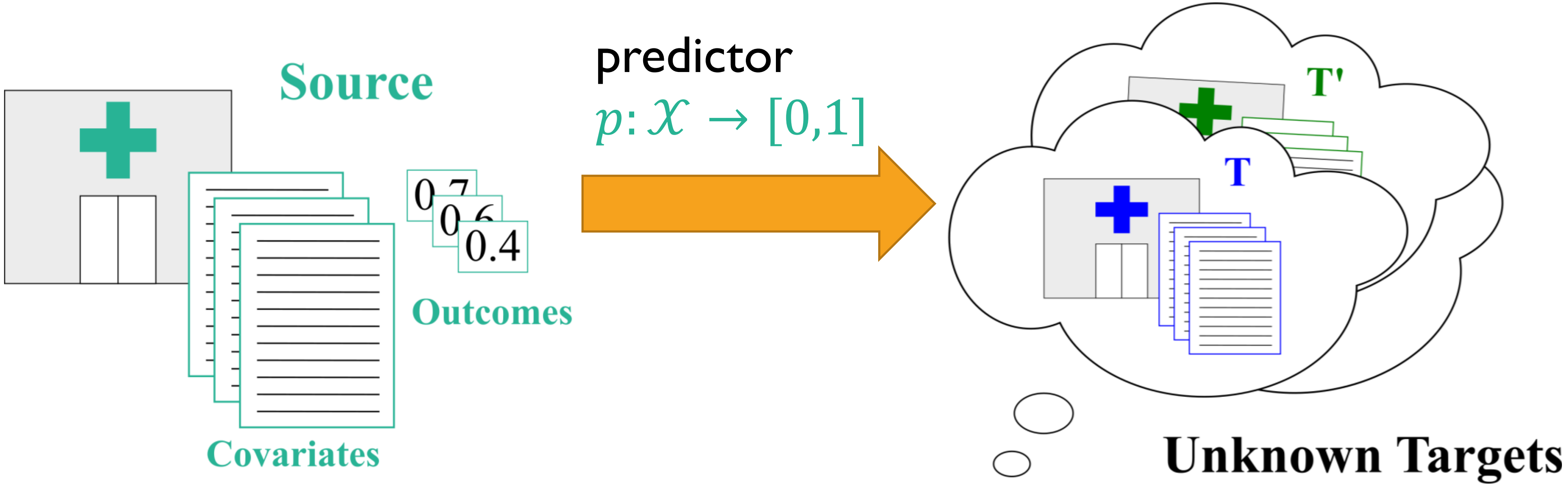
FITTING THE PROPENSITY SCORE

- Let Σ be a class of $\mathcal{X} \rightarrow [0,1]$ functions
  - Bounded complexity
- Fit an estimate σ of the propensity score $e_{st}(x)$
- Minimizing some loss function (say logistic loss)

CHALLENGES WITH THIS PARADIGM

- Assumptions may not always hold (we can't help this either).
- True propensity score may be far from Σ (we may have a chance).
- Need unlabeled samples from target for training. Not always feasible:
  - Apply the Stanford study to numerous other hospitals.
  - Apply the Stanford study to Stanford population in 5 years.
  - Limitations on sharing information.
- Universal adaptability?

UNIVERSAL ADAPTABILITY



UNIVERSAL ADAPTABILITY

- Train a predictor $p: \mathcal{X} \rightarrow [0,1]$ in source.
  - Labelled examples in source for training.
- Infer $E[Y|Z = t]$ in destination as $E[p(X)|Z = t\,]$.
  - Unlabeled examples in target for inference.
- Unfortunately: with standard loss minimization - predictor trained on source may give bad predictions on target (one path for discrimination).
- Can it be obtained?

ALGORITHMIC FAIRNESS

- Identifying forms of unfair discrimination and ways to address them.
- Historic oppression can manifest itself in Data: under-representation, mislabeling, missing features.
  - When can unfair discrimination through data be addressed?
- In heterogeneous populations, sometimes, notions of fairness can promote accuracy/utility (as it helps identify untapped potential and because inaccuracy can be a form of discrimination).
- This is such a story …

MULTICALIBRATION $\Rightarrow$ UNIVERSAL ADAPTABILITY

- Multicalibration [Hébert-Johnson, Kim, Reingold, Rothblum 18] developed and studied in the context of algorithmic fairness.
- Requires accurate (calibrated) predictions, not just overall, but on a large family of (large) subpopulations.
- Intuition: if estimator learned in source is multicalibrated it will directly apply for a target that weighs those subpopulations differently.

MULTICALIBRATION $\Rightarrow$ UNIVERSAL ADAPTABILITY

- For a class of functions $C \subseteq \{c: \mathcal{X} \to \mathbb{R}^+\}$, a predictor $\tilde{p}$ is (C, α)-multiaccurate, if for every $c \in C$
$$E\left[c(X) \cdot (Y - \tilde{p}(X))\right] \leq \alpha$$
- For a class of propensity scores Σ let $C(\Sigma) = \left\{\frac{1-\sigma(x)}{\sigma(x)} : \sigma \in \Sigma\right\}$
- Theorem: If p is $(C(\Sigma), \alpha)$-multiaccurate over source s, then p is (Σ, β)-universally adaptable for $\beta \leq \alpha + \delta_{st}(\Sigma)$.

EXTENSIONS AND EXPERIMENTS

- If p is $(C(\Sigma), \alpha)$-multiaccurate over source can infer $E[Y|Z = t]$ in target.
- If p is (C', α)-multicalibrated over source for larger class C' can infer more sophisticated statistics in target (p is (C'', α)-multicalibrated over target).
- Promising experiments – universal adaptability shows competitive and at times better performance than propensity scoring.
  - Intuitive when the propensity scores not in Σ

PROPENSITY SCORING FOR TREATMENT EFFECT

- Example: effect of vaccination on sever sickness.
- Usage 1: Adapt a study in source to a target.
  - Universal adaptability easily extends.
- Usage 2: Distribution of individuals with intervention ≠ distribution of individuals without intervention.
  - Propensity scoring can translate one to the other
  - So can multicalibration (requires some thought)
  - Multicalibration can also be used to learn propensity scores with subgroup guarantees [Gopalan, Reingold, Sharan, Wieder 22]

MUSINGS ON CAUSALITY AND MULTICALIBRATION

DISCRIMINATION BY NON-CAUSALITY

- A typical recipe of discrimination:
  - Select a small set of features that are highly correlated with outcome.
  - Fit a decision rule based on these features.
  - Fear: relation is non-causal and variables are proxies for protected attributes.
- Multicalibration allows taking into account a huge number of features and potential decision rules and simultaneously respecting them all
  - If some of these relations uncover causal relations, can we be happy?

MULTI-MODEL CAUSALITY?

- Since the introduction of multicalibration – an explosion of research (more than I can discuss here):
  - Additional fairness applications, related notions, additional algorithms
  - Applications beyond fairness

MULTI-MODEL CAUSALITY?

- Multicalibration gives an alternative to loss-minimization. Instead of optimization – indistinguishability:
  - Outcome indistinguishability [Dwork, Kim, Reingold, Rothblum, Yona21]
  - Universal Adaptability – no matter what the propensity score function is (within a class)
  - Omnipredictors [Gopalan, Kalai, Reingold, Sharan, Wieder 21] – minimization (compared to a class), no matter what the loss function we care about is.
- Can we have solutions that work no matter the causal model?

Empirical Evaluation

- Setting:
  - source US National Health and Nutrition Examination Survey
  - target US National Health Interview Survey
  - estimate 15-year mortality rate across demographic groups
- Results:
  - Imputation with a single multicalibrated predictor
  - Similar performance as demographic-specific PS estimates

Empirical Evaluation

- Semi-synthetic setting: simulate extreme covariate shift