Part 2 – Intervention and Causal Inference

- Understanding the effect of an intervention can address many hard problems in IR system
  - *Consequence* of changing algorithm, data pipeline, webpage design, ...
  - *Knowledge* about how users make decision (mechanism of the environment)
  - *Long-term utility / fairness* of our decision

- Standard statistical models no longer satisfy this purpose, because:
  - Intervention can be hypothetical and violating the natural course of observed data
  - Intervention can create alternative interpretations that may or may not be captured by regular rules, e.g. by conditional probability.

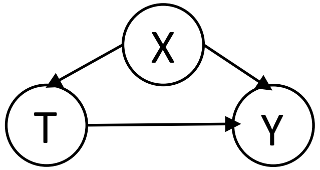
- The language of *causal inference* fills in the gap
  - Significantly emphasizes intervention within existing probability framework.

Pearl & Rubin causal model

- Recall that we wish to characterize everything related to *making intervention.* The solution from *Pearl's structural causal model*:
  - *Joint distribution* of the data, generated from the basic *noise variable* $\{U_i\}_{i=1}^{d}$
  - A *collection of equations* that formalize the *assumptions* of how the variables interact, e.g.

$$X_i := f_i(Pa(X_i), U_i), i = 1, \ldots, d$$

  - A *graphical model* that represent the assignment structure



  - *Assigning* values to certain variables specify a *response function,* via *do-operation*

$$p(Y \,|\, do(T := t)) \quad \text{(different from } p(Y \,|\, T = t))$$

  - *Average causal effect* of an intervention -> difference in the substitutions to the assignment:

$$\mathbb{E}[Y \,|\, do(T := 1)] - \mathbb{E}[Y \,|\, do(T := 0)]$$

Pearl & Rubin causal model

- From *conditional statement* to *interventional statement*
  - The biggest disagreement occurs with ***confounding*** – doing X=x may change something else and fail to coincide with conditional probability
  - But we can *control for* the confounding factors (marginalization):

$$p(Y = y \,|\, do(T = t)) = \sum_{x \in \mathcal{X}} p(Y = y \,|\, T = t, Pa(T) = x) \cdot p(Pa(T) = x)$$

  - The above *adjustment formula* allows us to estimate average causal effect from data. What about causal (***counterfactual***) questions other than the causal effect?
    - Observed evidence -> propagate the evidence to update the posterior of exogenous variables, e.g. $p(Pa(T) = x) \to p'(Pa(T) = x)$
    - Perform do-operation as usual with the updated distributions

Pearl & Rubin causal model

- If the assignment is *randomized* and the *intervention* takes the form of (binary) *treatment*, Rubin's model focus on the *potential outcome*
  - With n *units*, $(Y_1(i), Y_0(i)),\ i = 1, \ldots, n$ give the outcome under treatment/o.w.
  - It reflects the effect of intervention (treatment) more directly --

$$T(i) \in \{0,1\} \text{ as boolean treatment indicator, then } Y(i) = T(i)Y_1(i) + (1 - T(i))Y_0(i)$$

  - *Average causal effect* can be straightforwardly estimated, although we can only observe one potential outcome – suppose coin-toss assignment:

$$\mathbb{E}[Y(i)\,|\,T(i) = 1] = Y_1(i), \quad \mathbb{E}[Y(i)\,|\,T(i) = 0] = Y_0(i)$$

  - Let $\bar{Y}_1, \bar{Y}_0$ be the population average of $(Y_1(i), Y_0(i)),\ i = 1, \ldots, n$ then *average causal effect* = $\bar{Y}_1 - \bar{Y}_0$, because

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} Y(i),\,|\,T(i) = t\Big] = \bar{Y}_t \quad t \in \{0,1\}$$

The causal inference languages

- Despite the many conceptual debates, both models are useful for IR sys:
  - What variables can be intervened?
  - Is it possible to observe all confounding variables?
  - Reliability and usefulness of the structures among variables?
  - Source of randomness?
  - ...
  - A/B testing, offline studies, explainable IR, bias / fairness, ...

- *Causality and intervention* beyond average causal effect:
  - Which direction – $p(X,Y) = p(Y \mid X)p(X) \quad or \quad p(X,Y) = p(X \mid Y)p(Y)$
  - Ohm's law, altitude & temperature
  - *Intervention -> Invariance -> Independence & Causality*
  - This view will be useful for pattern recognition (later)

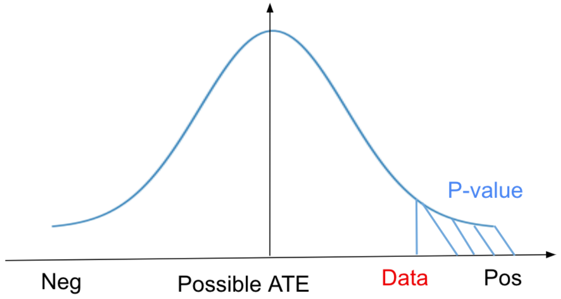
Design and inference

- How do we test whether an *intervention* can achieve desired outcome?
  - To exclude bias from all potential confounding, we design coin-toss assignment (as in Rubin's model and compute ***average causal effect***
  - If it is positive, is it positive just by chance? (*inference*: draw conclusion under uncertainty)
  - We shall use a *stochastic proof by contradiction*:

$$H_0:\ \text{non-positive vs. } H_1:\ \text{positive. How about } p(\text{more positive than what the data tells?} \,|\, H_0 \text{ is true})?$$ *

  - Intuitively, we can access $p(\text{some function of data} \,|\, H_0 \text{ is true})$ -- because of the design.

- *Hypothesis testing* and p-value
  - The above example is an instance of hypothesis testing
  - Central to the inference is ***p-value***



*: we use this type off non-rigorous notation for the sake of space.

Design and inference

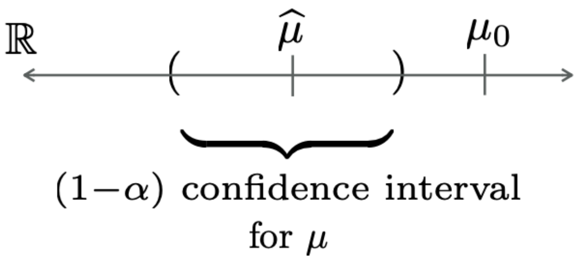
- *Hypothesis testing, p-value, significance level, and confidence region*:
  - What criteria to use? Reject H_0 if p-val < α – then p(false positive) < α!
  - Significance level α
  - Taking a detailed look at p-val < α for *average treatment effect* $Z = \bar{Y}_1 - \bar{Y}_0$

    Suppose the potential outcome are 1-subgaussian, equal sample in the *treatment group* and *control group. U*sing previous concentration result:

$$p\left(Z \leq \sqrt{\frac{2\log(1/\alpha)}{n}}\right) \leq \alpha \leftrightarrow \text{ reject } H_0 \text{ when } 0 \notin \left[\hat{z} - \frac{2\log(1/\alpha)}{n}, \infty\right]$$

  - There is an equivalence between rejecting based on *p-value* and *confidence region* for $H_0 : \mu = \mu_0$

$$\text{p-value} \leq \alpha \leftrightarrow \mu_0 \notin [LCB(n_1, n_2, \alpha), UCB(n_1, n_2, \alpha)]$$



where LCB, UCB are the lower/upper confidence bound.

A/B testing, metric, continuous monitoring

- When comparing two systems online, users are randomly bucketed and assigned to experience each system (A/B testing). Practically:
  - If there is a performance difference, we hope to detect it / reject the null hypothesis asap.
  - P-value is constantly checked to monitor the progress.

- The *sensitivity* of metric
  - For IR sys, online testing metrics have more room to explore
  - Reduce the variance of a single metric, or combine multiple metrics smartly
  - ***Rao-Blackwell Theorem***: using sufficient statistics to construct metric with smaller variance

$$\mathbb{E}[(\theta - \mathbb{E}[\theta \,|\, T])^2] \leq \mathbb{E}[(\theta - \hat{\theta})^2], \text{ for all } \hat{\theta}$$

  - ***Linear Discriminant Analysis***: linear combination of metrics that optimizes Z-score

$$\max_{\theta} \frac{\bar{Z}_1 - \bar{Z}_0}{\sqrt{var(\bar{Z}_1 - \bar{Z}_0)}} \quad s.t. \quad Z = \theta^T [Y^{(1)}, \ldots, Y^{(d)}]$$

A/B testing, metric, continuous monitoring

- Recall that *significance level* α is designed for *one-time control* of false positive rate *under fixed sample size*
  - Let P$^{(n)}$ be the p-value obtained from the *first-n* samples
  - Under null hypothesis, given a fixed n, it holds $p(P^{(n)} \leq \alpha) \leq \alpha$ (***uniformity***)
  - In *continuous monitoring*, the test is continued if p-val > α, so the real *stopping time is*

$$\tau := \min\{n \in \mathbb{N} : P^{(n)} \leq \alpha\}$$

  - Note that $p(P^{\tau} \leq \alpha)$ can be much bigger than α. (why?) False positive becomes very likely!
- How to make sure p-value is *always valid*, e.g. satisfy uniformity?
  - Recall that $\text{p-value} \leq \alpha \leftrightarrow \mu_0 \notin [LCB(n_1, n_2, \alpha), UCB(n_1, n_2, \alpha)]$
  - The previous confidence regions are derived for the average of i.i.d variables
  - Under continuous monitoring, $Y_1, \ldots, Y_n$ are dependent, so $\bar{Y}$ is a random walk.
  - Using concentration bound for random walk!

Continuous monitoring, best-arm identification

- Develop confidence regions for random walks -> *always valid* p-value:
  - Confidence band for i.i.d average: $\mathcal{O}(\sqrt{\frac{log(1/\alpha)}{n}})$
  - Confidence band for random walk average: $\mathcal{O}(\sqrt{\frac{\log\log n + C\log(1/\alpha)}{n}})$ [ZZS+16]
  - Achieves $p(\bigcup_{n\in\mathbb{N}} \mu_0 \notin [LCB(n,\alpha), UCB(n,\alpha)]) \leq \alpha$ under null hypothesis $H_0 : \mu = \mu_0$
- By making p-value any-time, we have more choices with adaptive testing:
  - *Adaptively* update the pool of candidates, without sacrificing the rigor of testing
  - Be smarter in traffic directing – good candidates deserves more samples, while exploring bad candidates just enough to safely eliminate / replace them
  - ***Best-arm identification*** algorithm that also relies on upper/lower confidence bounds

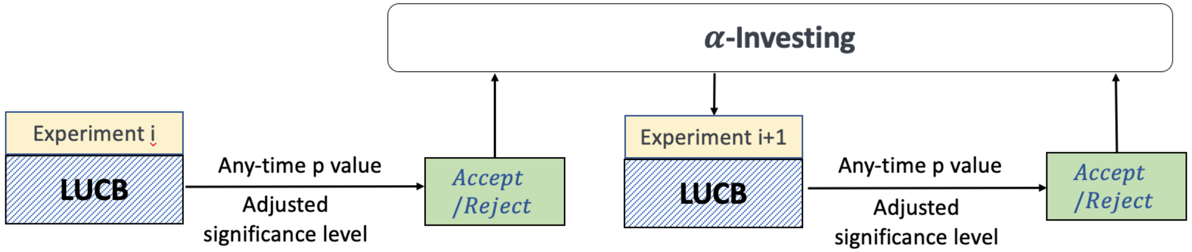
Best-arm identification, sequential testing

- Best-arm identification with pure exploration methods:
    - Always-valid p-value allows us be more adaptive (creative) in finding the best candidate system
    - ***LUCB*** method for pure exploration
        1. obtain equal amount of initial feedback for each system
        2. keep the traffic to the *current-best*, *second-best* and the *control system* (a^*, a^{**}, a^0)
        3. Update a^*, a^{**}, and iterate with step 2 until: [*LCB of* a^*] > [*UCB of* a^{**} *and* a^0]

- Integrate testing with LUCB
    - Can we just compute the always-valid p-value when LUCB stops, and decide how to proceed?
    - If we do this multiple times, the significance level α no longer guarantees the *online false discovery rate (FDR)* of the sequence of tests. (why?)
    - The significance level also needs to be updated every time an *accept/reject decision* is made
    - **α-investing** method for online FDR control: "invest" (discount) α each time when a testing is called, "reward" (increase) α when making discovery. [FS08]

Best-arm identification, sequential testing

- What cause the inherent difficulty of the algorithm? It has been shown with a information-theoretical lower bound that:
  - How “spread-out” the gaps are
  - The gap between the best and second-best arm

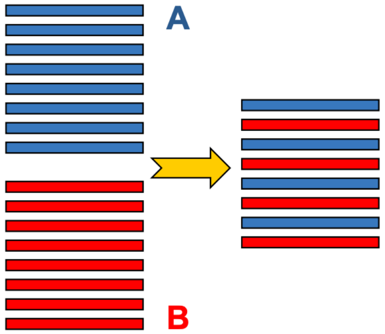


  - power? adding context? ....
- Some fundamental issues for A/B testing
  - absolute feedback, venerable to *between-subject* variability (by how much?)
  - when comparing ranking outcomes in particular, design *within-subject* relative comparison?

Interleaving and dueling

- Interleaving -- eliminate noise by letting user compare both alternatives
  - more robust to users' decision bias
  - less affect user experience
  - clicks directly reflect users' preference for A vs. B

[YJ09]



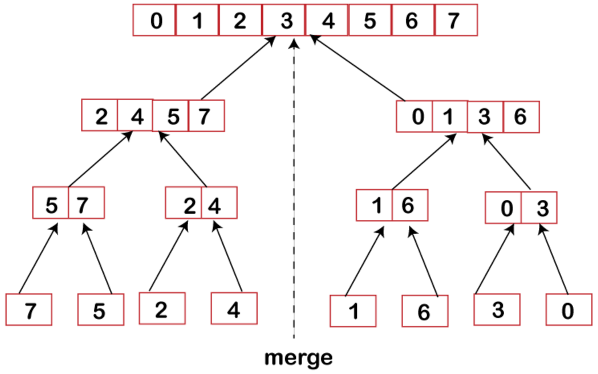
- Set up interleaved outputs – again, room to optimize *sensitivity*
  - *Balanced interleaving, Team Draft, Probabilistic Interleave*
  - A ***probability distribution*** to show a particular combination (think about randomized treatment assignment)
  - A ***scoring rule*** to interpret click -- a measure for treatment effect, $H_0 : \text{score}_A = \text{score}_B$

- *Optimization* via random user model and max-entropy principle:
  - ***Optimization*** *variable*: probability to show each page
  - ***Constraint***: a model of random user, express no preference
  - ***Objective***: maximizing the *entropy* (uncertainty of having a winner)

Interleaving and dueling

- Making interleaving test *adaptive* with K ranking systems
  - "Dueling" – create a schedule for *pairwise comparison* to find the best candidate
  - The idea of *using lower/upper confidence bounds* to compare under uncertainty is still valid
  - **Key challenge**: pairwise comparison to determine total order with K systems, $\mathcal{O}(K^2)$?
  - *Example: interleaved filter*
    - *randomly pick a^* to compare with all others*
    - *repeat until finding a^{**} whose LCB/UCB goes beyond those of a^**
    - *elimination, repeat*

- Integrating with algorithms from ***sorting***
  - Reduce the number of dueling needed to determine the order
  - Achieves $\mathcal{O}(K^2) \rightarrow \mathcal{O}(K)$
  - How many samples are needed? (topics for later)
  - Compatible with adaptive online testing? (reduction to cardinal bandit [AKJ14])



Thinking about the assumptions

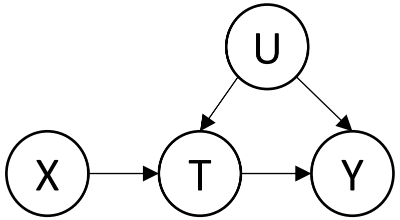
- Acute audience may find a critical background missing – what makes *randomized intervention* work for causal inference in the first place?
  - ***Stable Unit Treatment Value Assumption (SUTVA)***

    treatment one unit receives does not change the effect for another unit
  - ***Consistency***

    true outcome agrees with the potential outcome given the treatment indicator
  - ***Ignorability***

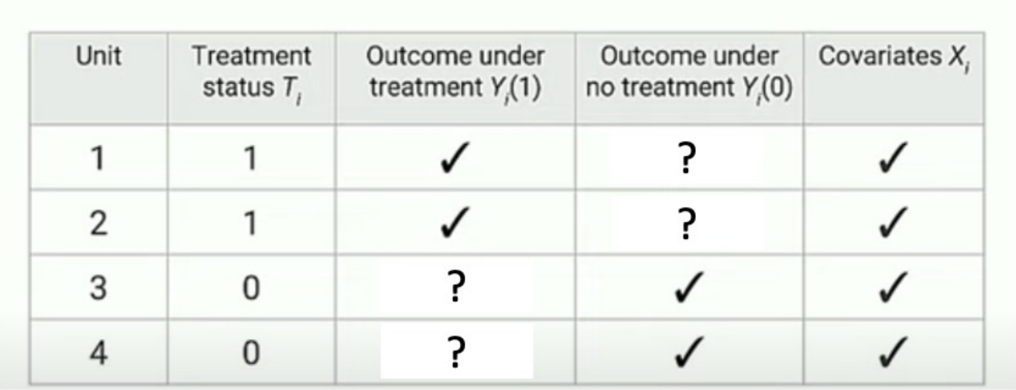
    potential outcome conditionally independent of treatment given defounding variables

- Unfortunately, they are all violated to a degree in IR sys...
  - Spillover, network and equilibrium effect
  - Leap between *exposure as measured* and *exposure as intervened*?
  - In the presence of unobserved confounding, are potential outcome ***missing-at-random (MAR)***? More importantly, can causal inference problem be treated as a missing data problem?



Observational studies and offline learning

- When the ability to launch randomized intervention is limited, or would like to mine the logged data from experimentation
  - Does the problem reduce to *pattern recognition with feedback data*?
  - Incorporate *causal knowledge* to address the *partial observability* of potential outcome?
  - Problem solved if we estimate "?" →
  - Studying causal problem as missing data problem attracts huge attention in IR



Unit	Treatment status T_i	Outcome under treatment $Y_i(1)$	Outcome under no treatment $Y_i(0)$	Covariates X_i
1	1	✓	?	✓
2	1	✓	?	✓
3	0	?	✓	✓
4	0	?	✓	✓

- Unfortunately, these two aren't the same [PM18]
  - Domain ***overlapping***?
  - ***Identifiability*** of causal mechanism (invariant mechanism)?
  - When ***imputing*** missing data require unsupported ***extrapolation ...***

Is observational studies a missing data problem?

- Admittedly, if the missing mechanism is simple enough, life becomes much easier:
  - If feedback is *missing completely at random* (**MCAR**), which means the ***missing mechanism*** – which we can consider as treatment, is assigned by coin toss
  - If feedback is *missing at random* (**MAR**), the missing mechanism is still random but may depend on some confounding factors
  - In both cases, *average causal effect* can be effectively estimated using previous techniques

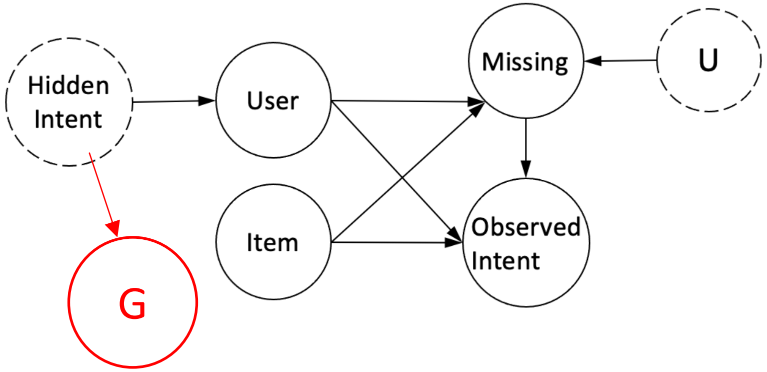
- When feedback is missing not at random (**MNAR**)
  - Can be caused by *missing with certainty,* e.g. some items have zero chance to be exposed
  - The crucial ***positivity*** assumption is violated, and positivity is needed to ensure ***overlap*** between treatment / control group
  - If a subgroup of subjects always receives the *same intervention*, we cannot estimate the effect of intervention changes on that subgroup *without further assumptions*
  - What further assumptions are needed? – e.g ***identifiability*** of certain causal pathways [MP21]

Is observational studies a missing data problem?

- In IR system, user selection bias -> *hidden intent* often causes the missingness of the *observable intent*
- "Self-masking" – extremely challenging
- How to use survey data to estimate the average income of low-income

  family when they family don't have money to install phone and answer the survey?

  -- Nobel-winning solution in Economics, *Heckman correction*



- Generally impossible to impute the missing feedback (unobserved intent) with guarantee, unless:
  - The pathway *{U -> missingness}* is known and satisfy *positivity* (e.g. fully randomized recommendations with known policy)
  - *Side information* about the hidden intent, e.g. hidden intent causes the clustering geometry of users (Node G), which is often observed in IR data
  - The pathway of *{hidden intent -> G}* , which is an *independent mechanism*, will assist extrapolation as we discuss later.

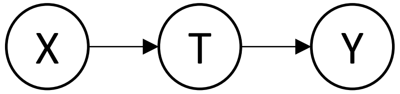
[XY22]

Double learning and targeted maximum likelihood learning

- Can we leverage modern ML methods as estimators, and plug their predictions for estimating average causal effect ψ?
  - Recall the *adjustment formula*

$$p(Y = y \mid do(T = t)) = \sum_{x \in \mathcal{X}} p(Y = y \mid T = t, Pa(T) = x) \cdot p(Pa(T) = x)$$

  - Let $X = Pa(T)$ -- use neural network to obtain $\hat{p}(Y \mid T, X)$
  - Big issue: most *finite-sample* ML-based estimator are biased! (e.g. the use of regularization)
    confidence interval obtained in the usual way may not cover the true average causal effect!
  - How about $\hat{\psi}^Q := 1/n \sum_{i=1}^{n} (\hat{Q}(1, x) - \hat{Q}(0, x)), \quad \hat{Q}(t, x) = \mathbb{E}[Y \mid t, x]$?

-- not using 

: X affect Y only through treat assignment $g(X) := p(T = 1 | X = x)$

  - Sufficiency of ***Propensity Score*** [RR83]:

$$\psi = \mathbb{E}\big[\mathbb{E}[Y \mid g(X), T = 1] - \mathbb{E}[Y \mid g(X), T = 0]\big]$$

Double learning and targeted maximum likelihood learning

- *Semi-parametric estimation theory* to the rescue under *unknown* propensity scores
  - Recall that we estimate $\hat{Q}(x,t) = \mathbb{E}[Y \mid X = x, T = t]$, $\hat{g}(x) = p(T = 1 \mid X = x)$
  - These estimation from ML models could be *biased*, hurting the *asymptotic properties* (e.g. confidence interval may not cover the true average causal effect ψ)
  - Fortunately, $\hat{\psi}$ can still have good asymptotic property if satisfies [Ken16]:

$$\frac{1}{n}\sum_i \varphi(y_i, t_i, x_i; \hat{Q}, \hat{g}, \hat{\psi}) = 0$$

  - "*Non-parametric estimating equation*" with "*efficient influence curve*" (think of first-order bias under Taylor expansion)
  - ***Targeted maximum likelihood learning*** [VR06] – solve the estimating equation by perturbing $\hat{Q}$ using some parametric submodel, e.g. $\hat{Q}^{(1)} = \hat{Q} + \epsilon H(\hat{g})$, H is given, and use MLE to estimate ϵ
  - ***Double/debiased ML*** [CCD+18]*:* use a ***Neyman-orthogonal score*** equation for first-order debias, computing a cross-fitted augmented IPW estimator
  - As long as $\hat{Q}$ and $\hat{g}$ are faster than $n^{1/4}$, then $\hat{\psi}$ enjoys $n^{1/2}$-rate *asymptotic normality*

Propensity weighting method and counterfactual learning

- When treatments are characterized by *known* distributions, address changing the treatment on a population as switching the distribution
  - Easily verify mathematically that: $\mathbb{E}[Y \mid do(T=1)] = \mathbb{E}\big[YT \,/\, \mathbb{E}[T=1|X=x]\big]$
  - The many causal assumptions, especially positivity, ensures the *unbiasedness* of the estimation:

$$\mathbb{E}[Y \mid do(T=1)] - \mathbb{E}[Y \mid do(T=0)] = \mathbb{E}\Big[Y\Big(\frac{T}{\pi(X)} - \frac{1-T}{1-\pi(X)}\Big)\Big],\ \pi(X) = p(T=1 \mid X=x)$$

  - What about the variance? (often need *strong positivity / overlapping*)
- If we have a good estimation for the *average potential reward* using collected data, can we use that estimation to learn a good policy?
  - ***Counterfactual learning*** : $\arg\max_{\pi_\theta} \mathbb{E}[Y \mid do(T=\pi_\theta(X))]$
  - Moving from treatment to action $a \in [K]$. Propensity score thus becomes: $\pi_0(A=a_i \mid X=x_i)$

so we have: $\hat{R}(\pi_\theta) := \hat{\mathbb{E}}[Y \mid do(\pi_\theta(A|X))] = \dfrac{1}{n}\sum_i y_i \dfrac{\pi_\theta(a_i|x_i)}{p_i}, \quad p_i = \pi_0(a_i \mid x_i)$

Propensity weighting method and counterfactual learning

- Is counterfactual learning a supervised learning problem with weighted loss?
  - Not exactly. The observations are different: $(x_i, \text{instructive } y_i^*)$ versus $(x_i, a_i, p_i, \text{evaluative } y_i)$
  - Access to the loss are different: $\ell(y_i^*, y)$ known, versus $\ell(y, f(x_i))$ unknown for $y \neq y_i$
  - Despite the conceptual difference, what makes counterfactual learning difficult?

hypothesis + optimization

- *Distribution shift* and *variance of risk estimator* for hypothesis
  - We've seen before in LTR the *Bernstein-type bound* – variance of risk estimation matters!
  - The variance is going to be large with small probability on denominator ...

$$\mathcal{O}\Big(\sqrt{\frac{V_n(f)\log(\mathcal{C}(\mathcal{F})/\delta)}{n}}\Big)$$

  - ***Clip*** small probability, ***renormalize*** the propensities, or ***penalize the variance*** in general

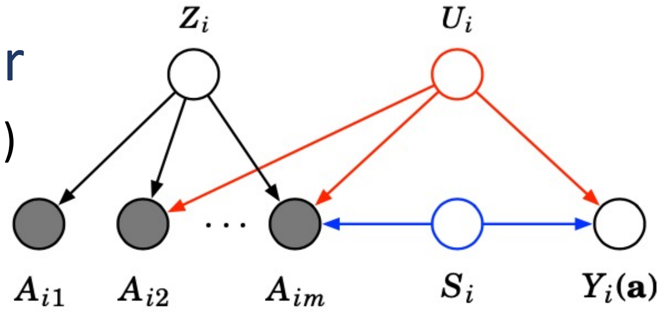
- Importance of having a *baseline* for optimization
  - If all feedback are non-positive, what will happen? (*degeneracy* of the learning problem because a upper bound can be trivially found)
  - Find a good $y_i' = y_i - r(a_i, x_i)$ to make the learning and optimization more *robust*.

[SJ15]

Multiple causes, deconfounding, robust optimization

- We talked about unobserved confounding makes inference from observational data infeasible
  - may be compatible with many potentially contradictory causal explanations
  - how much information about unobserved confounding can be recovered from observed data?

- Infer a latent variable as a substitute for unobserved confounder
  - Suppose there are multiple causes (e.g. *MF factor models* in IR [WLC+18])
  - Assume *SUTVA, overlap*, and some parametric forms
  - If we can find such a *proxy* Z, then we can safely ignore U [WB19]
  - Can employ some *encoder-decoder* learning framework (trade assumptions for assumptions)



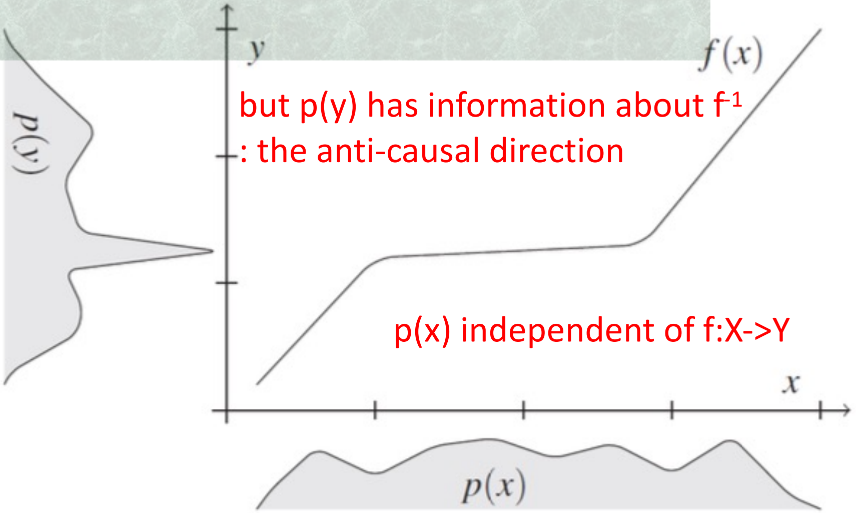
- How sensitive are the outcome?
  - Adding ad-hoc violations to the causal assumptions, and investigate the resulting perturbations
  - Or making the learning/optimization *robust to the violation* of causal assumptions [XRK+20]

Connection to IR pattern recognition

- Causality and learning (plenty in IR)
  - Predict target from its *cause* vs. from its *effect*
  - The principle of ***independent causal mechanism***

    [PJS17] -- independent mechanism + autonomous modules,

    and they do not inform each other
  - Exploit the *independent mechanisms*, e.g. via *causal discovery* or *directional learning*, and use them to assist generalizing to unseen data



but p(y) has information about f^{-1}
: the anti-causal direction

p(x) independent of f:X->Y

- Uncertainty quantification, learn-then-test

$$p(Y_{\text{test}} \in S(X_{\text{test}})) \geq 1 - \alpha$$

  - Creating statistically rigorous prediction sets for ML predictions (IR cares *coverage*!)
  - ***Distribution-free conformal prediction*** – use quantiles of calibrated scores
  - Learn-then-test to optimize the *converge risk:* $\mathcal{R}(S_\lambda) := p(Y_{\text{test}} \notin S_\lambda(X_{\text{test}}))$

    suppose the *coverage set* S_λ depends on a parameter λ [AB22]
  - Hypothesis testing for whether the risk is controlled for a particularλ + *FDR control*

$$\{H_\lambda : \text{ the risk is controlled at } \lambda;\ \lambda \in \Lambda\} \Longrightarrow \mathrm{p}(\sup_{\lambda \in \Lambda} \leq \alpha) \geq 1 - \delta$$

Summary of Part 2

- Exploiting the intervenability of IR systems
  - Designed experiments -> answering causal questions
  - Problem structure + policy optimization -> gaining *efficiency* for IR experiments
- Counterfactual reasoning under domain practice
  - How valid are the *causal assumptions* under common IR practices?
  - May be more difficult than a *missing data problem*
  - Again, might need to incorporate *domain knowledge* for the rescue
- Offline pattern recognition with experimental feedback data
  - Knowledge about *intervention* makes learning from evaluative feedback favorable
  - Be aware of the variance caused by *distribution mismatch*
  - *Causality* can assist learning, but we may have to make deals with the devil of *confounding*