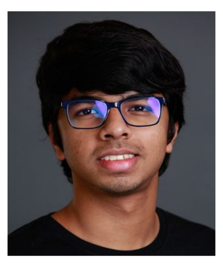
On Counterfactual Inference with Unobserved Confounding


Abhin Shah
abhin@mit.edu

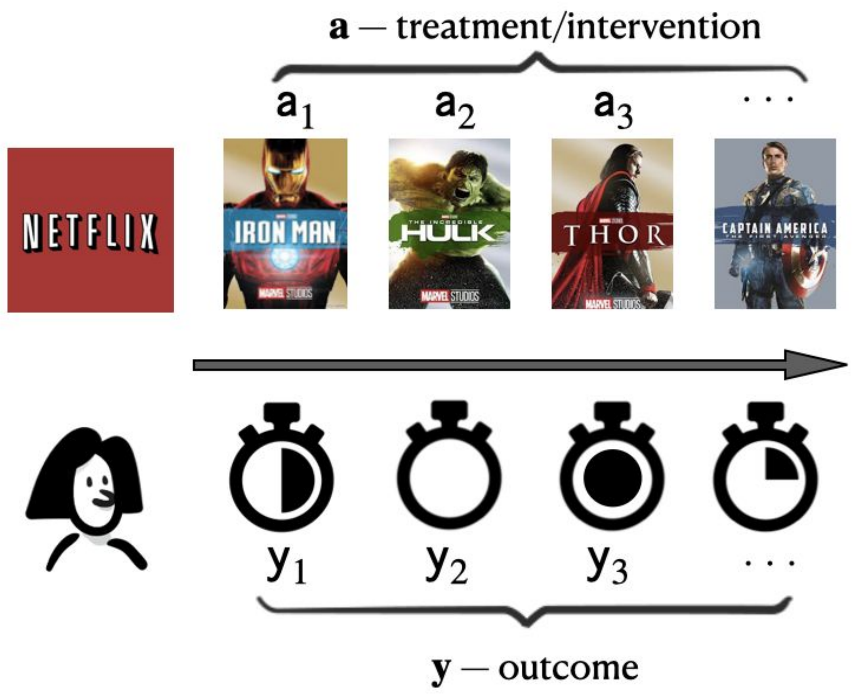

Raaz Dwivedi
raaz@mit.edu

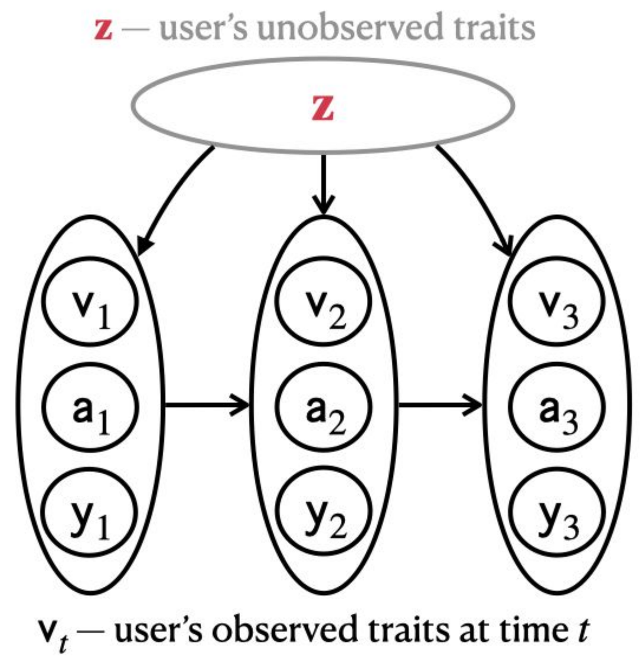

Devavrat Shah
devavrat@mit.edu


Greg Wornell
gww@mit.edu

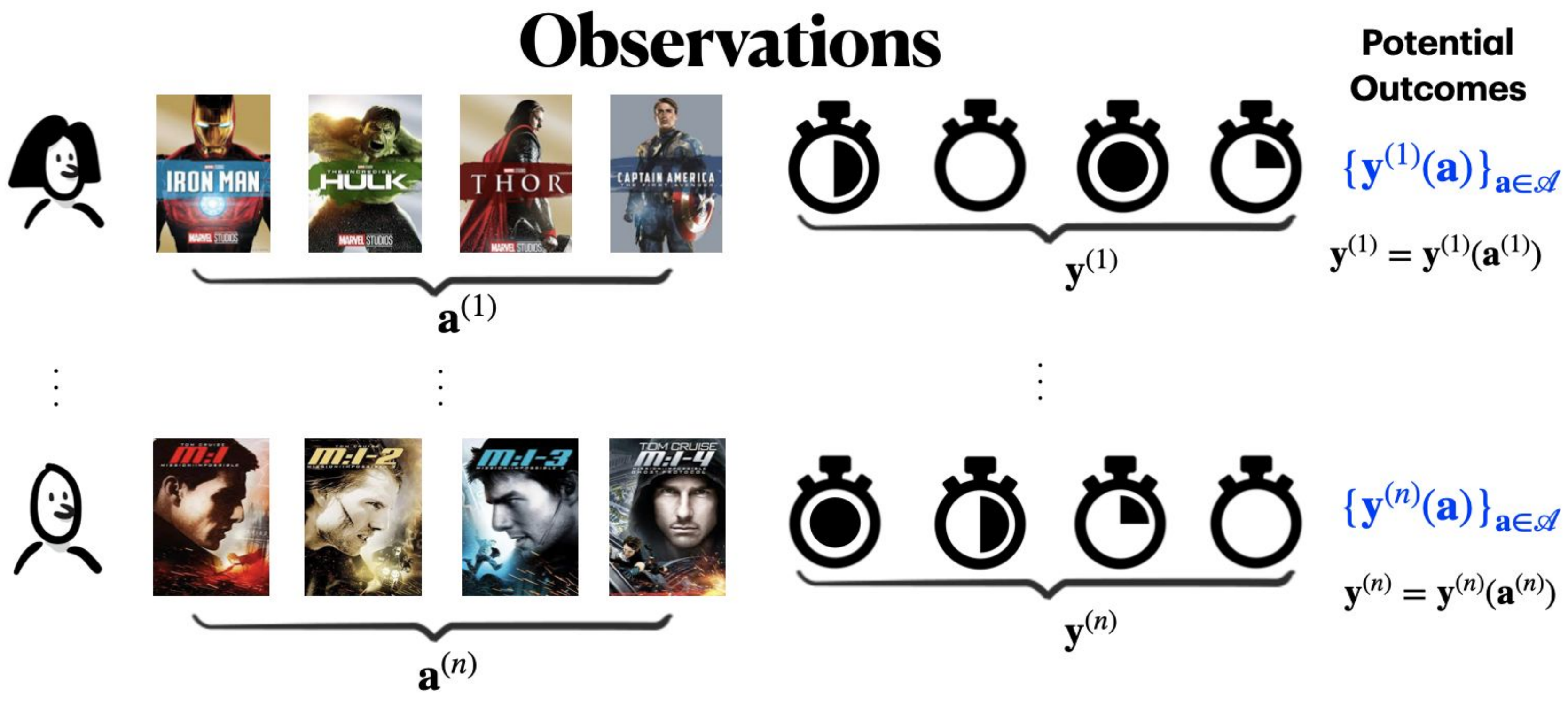
arxiv.org/pdf/2211.08209

Sequential Recommender System

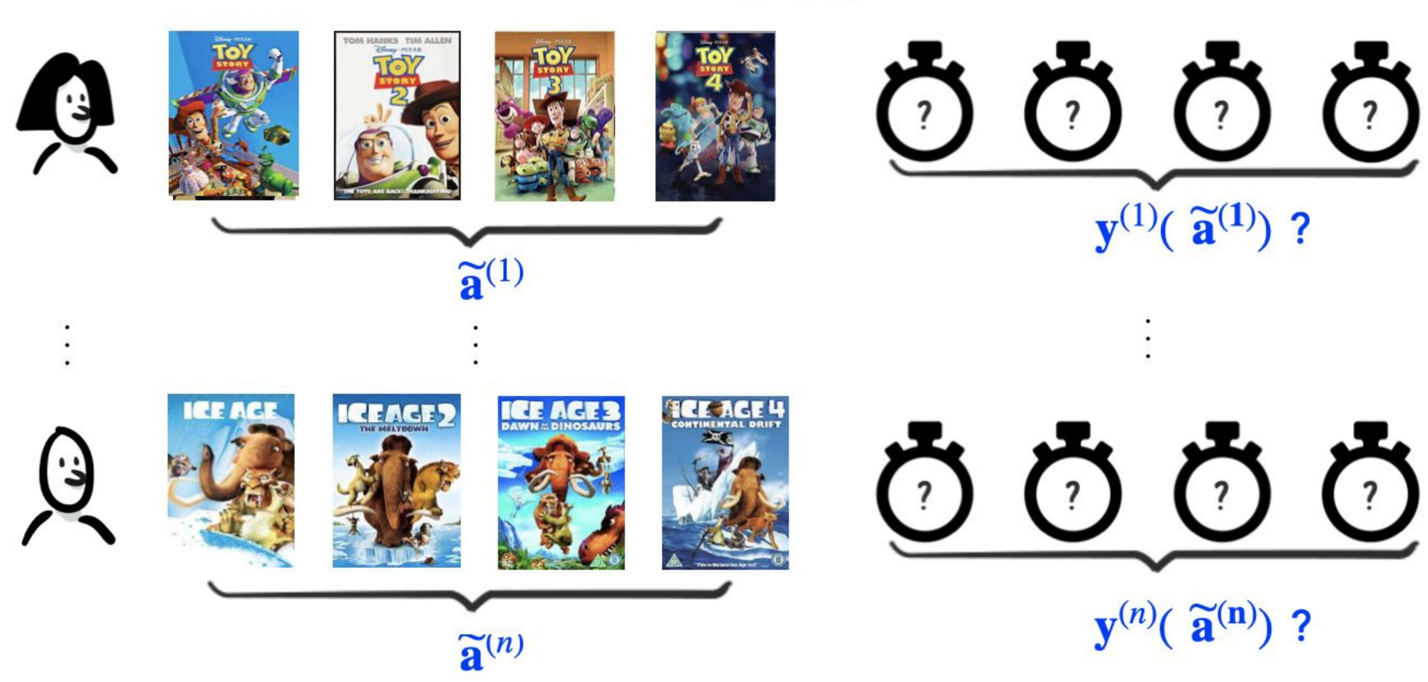




Observations

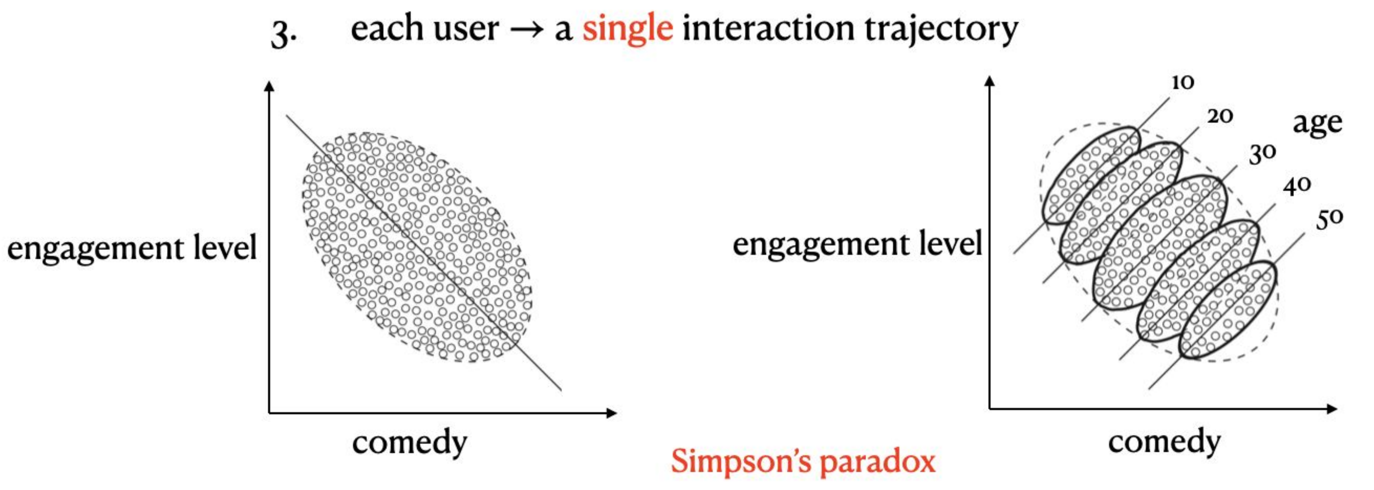


Goal

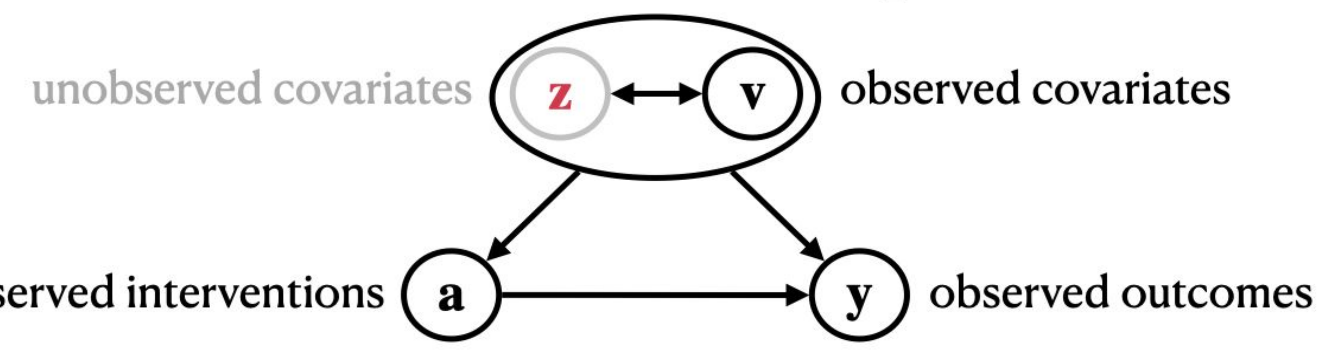


Challenges

1. unobserved factors → spurious associations
2. users → heterogeneous
3. each user → a single interaction trajectory



Problem Setup



n heterogenous and independent users with one observation each - $\{\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1}^n$

Goal : Counterfactual Questions

For user $i \in [n]$, what would have happened if alternative treatments were assigned?

$\equiv$

$\boxed{\text{Estimate } \boldsymbol{y}^{(i)}(\tilde{\boldsymbol{a}}^{(i)}) \text{ for } \tilde{\boldsymbol{a}}^{(i)} \in \mathcal{A}?}$

Suffices to learn $p(\mathbf{y} = \cdot \mid \mathbf{a} = \cdot, z^{(i)}, v^{(i)})$ for all $i \in [n]$, but each user may have *different z*

Can we learn n different distributions with *one* sample per distribution ?

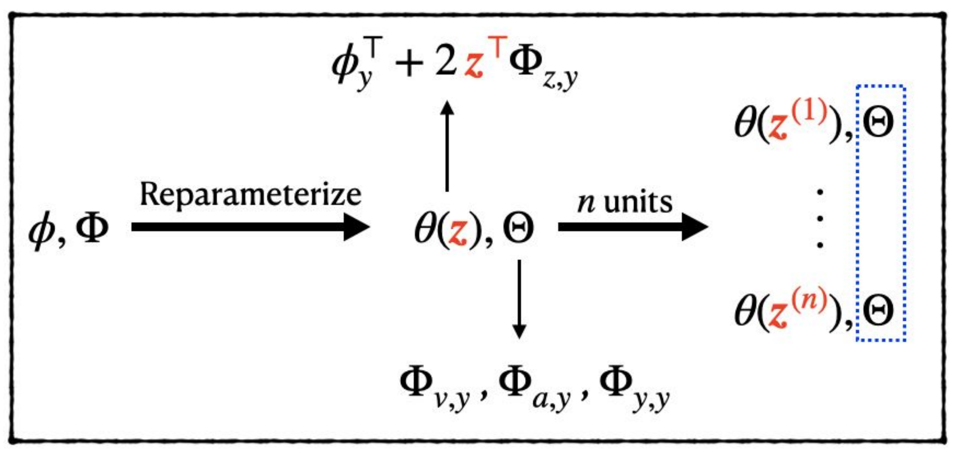
Our Approach

We posit a joint exponential family distribution for $\mathbf{w} \triangleq (\mathbf{z}, \mathbf{v}, \mathbf{a}, \mathbf{y})$

$$p(\boldsymbol{w}) \propto \exp\left(\boldsymbol{\phi}^\top \boldsymbol{w} + \boldsymbol{w}^\top \boldsymbol{\Phi} \boldsymbol{w}\right)$$

$$p(\boldsymbol{y} \mid \boldsymbol{a}, \boldsymbol{z} = \boldsymbol{z}^{(i)}, \boldsymbol{v} = \boldsymbol{v}^{(i)}) \propto \exp\Big(\Big[\underbrace{\boldsymbol{\phi}_y^\top + 2\boldsymbol{z}^{(i)\top}\boldsymbol{\Phi}_{z,y} + 2\boldsymbol{v}^{(i)\top}\boldsymbol{\Phi}_{v,y}}_{\text{different for different users}} + 2\boldsymbol{a}^\top\boldsymbol{\Phi}_{a,y}\Big]\boldsymbol{y} + \boldsymbol{y}^\top\boldsymbol{\Phi}_{y,y}\boldsymbol{y}\Big)$$

n heterogeneous conditional distributions ⇒ same exp. family but with diff. parameters



Inference Tasks

1. Parameters: User-level — $\theta^\star(z^{(i)})$ for all $i \in [n]$ → counterfactual distribution

Population-level — $\Theta^\star$

2. Potential Outcomes: $\mu^{(i)} \triangleq \mathbf{E}\left[\boldsymbol{y}^{(i)}(\tilde{\boldsymbol{a}}^{(i)}) \mid \mathbf{z} = z^{(i)}, \mathbf{v} = v^{(i)}\right]$ → counterfactual mean

Parameter Estimation

$\{\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1}^n$ → $\hat{\theta}^{(1)}, \cdots, \hat{\theta}^{(n)}, \hat{\Theta}$ estimates

pool all n samples

$$\min_{\theta^{(1)},\cdots,\theta^{(n)},\Theta} \sum_{t\in[\text{dim}]} \frac{1}{n}\sum_{i\in[n]} \exp\left(-\left[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}\right] x_t^{(i)}\right)$$

Assum 1: $\Theta^\star$ has sparse rows

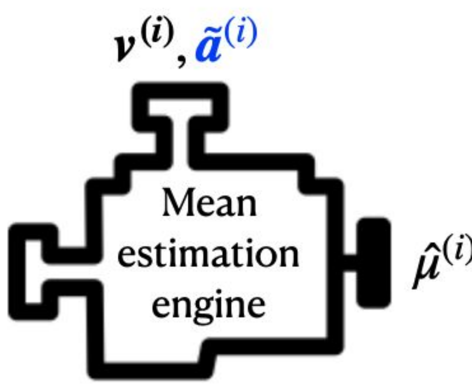
Assum 2: $\theta^\star(z^{(i)}) \in$ set $\mathscr{B}$

$\lVert\Theta^\star - \hat{\Theta}\rVert_{2,\infty} \le \epsilon$	when $n \ge O\left(\frac{\log(\text{dim})}{\epsilon^4}\right)$
For all i, $\lVert\theta^\star(z^{(i)}) - \hat{\theta}^{(i)}\rVert_2 \le \max\{\epsilon, M\}$	when $n \ge O\left(\frac{\text{dim}^2 M^2}{\epsilon^4}\right)$

metric entropy of $\mathscr{B}$

$M = O(s\log(k))$ when $\mathscr{B}$ = s-sparse linear combinations of k known vectors

Outcome Estimation



$$\hat{p}(\boldsymbol{y} \mid \boldsymbol{a} = \tilde{\boldsymbol{a}}^{(i)}, \boldsymbol{z} = \boldsymbol{z}^{(i)}, \boldsymbol{v} = \boldsymbol{v}^{(i)}) \propto \exp\Big(\Big[\hat{\theta}(\boldsymbol{z}^{(i)}) + 2\boldsymbol{v}^{(i)\top}\hat{\boldsymbol{\Phi}}_{v,y} + 2\tilde{\boldsymbol{a}}^{(i)\top}\hat{\boldsymbol{\Phi}}_{a,y}\Big]\boldsymbol{y} + \boldsymbol{y}^\top\hat{\boldsymbol{\Phi}}_{y,y}\boldsymbol{y}\Big)$$

For all i and any $\tilde{\boldsymbol{a}}^{(i)} \in \mathcal{A}$,

$$\boxed{MSE(\mu^{(i)}, \hat{\mu}^{(i)}) \le \frac{s\log(k\cdot\text{dim}) + \epsilon^2}{\text{dim}} \quad \text{when } n \ge O\left(\frac{s\,\text{dim}^2\log(k\cdot\text{dim})}{\epsilon^4}\right)}$$

Application: Denoise User-wise Covariates

No unobserved covariates

Noisy observed covariates = true covariates + measurement error

$\bar{\mathbf{v}} \quad = \quad \mathbf{v} \quad + \quad \Delta\mathbf{v}$

Assum 1: Only half users have error: $\Delta\mathbf{v}^{(i)} = \mathbf{0}$ for $i \in \{n/2, \cdots, n\}$

Assum 2: Covariates have sparse error: $\lVert\Delta\mathbf{v}^{(i)}\rVert_0 \le s$ for $i \in \{1, \cdots, n/2\}$

Goal: Estimate the true covariates

$$\boxed{\text{For all } i \in [n/2],\ MSE(\mathbf{v}^{(i)}, \hat{\mathbf{v}}^{(i)}) \le \frac{s\log(\text{dim})}{\text{dim}} + \epsilon^2 \quad \text{when } n \ge O\left(\frac{s\log(\text{dim})}{\epsilon^4}\right)}$$