# On Counterfactual Inference with Unobserved Confounding

Abhin Shah
MIT

Raaz Dwivedi
Harvard & MIT

Devavrat Shah
MIT

Greg Wornell
MIT

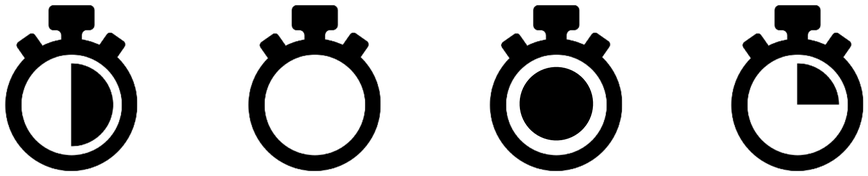
# Observational Setting


$a_1$ $a_2$ $a_3$ $\cdots$





**a** — treatment/intervention

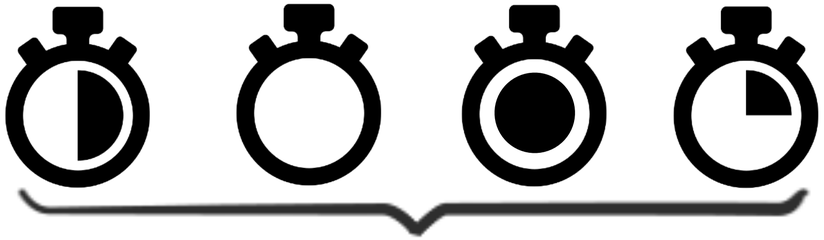


$y_1$ $y_2$ $y_3$ $\cdots$

**y** — outcome

# Panel Data


$\mathbf{a}^{(1)}$



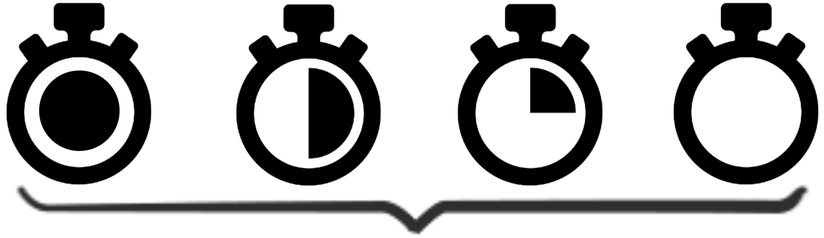
$\mathbf{y}^{(1)}$

**Potential Outcomes**

$\{\mathbf{y}^{(1)}(\mathbf{a})\}_{\mathbf{a}\in\mathscr{A}}$

$\mathbf{y}^{(1)} = \mathbf{y}^{(1)}(\mathbf{a}^{(1)})$


$\mathbf{a}^{(n)}$



$\mathbf{y}^{(n)}$

$\{\mathbf{y}^{(n)}(\mathbf{a})\}_{\mathbf{a}\in\mathscr{A}}$

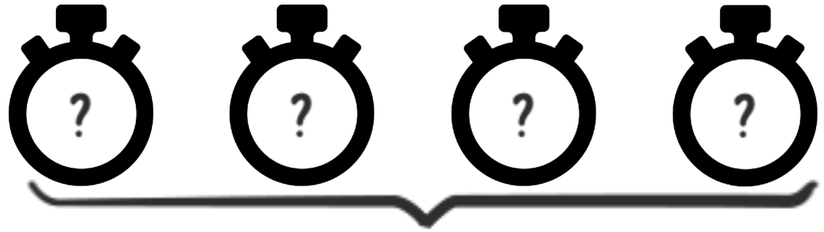
$\mathbf{y}^{(n)} = \mathbf{y}^{(n)}(\mathbf{a}^{(n)})$

# Goal



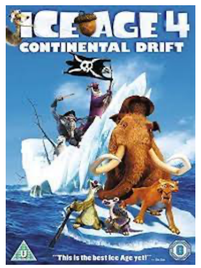


$\tilde{\mathbf{a}}^{(1)}$

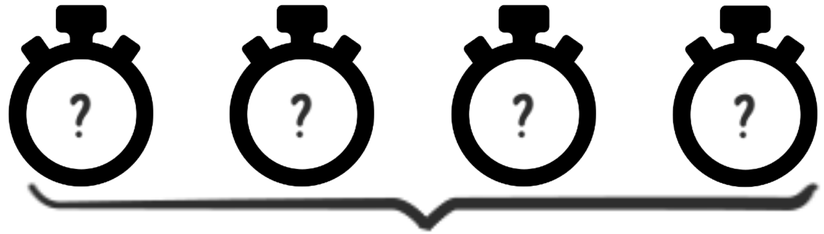


$\mathbf{y}^{(1)}(\tilde{\mathbf{a}}^{(1)})$ ?



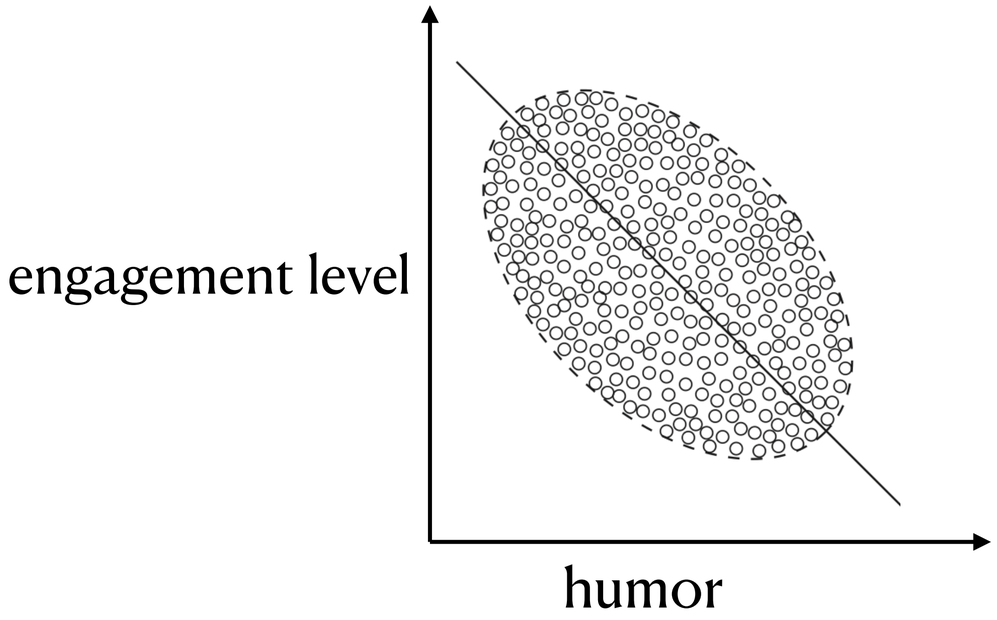

$\tilde{\mathbf{a}}^{(n)}$

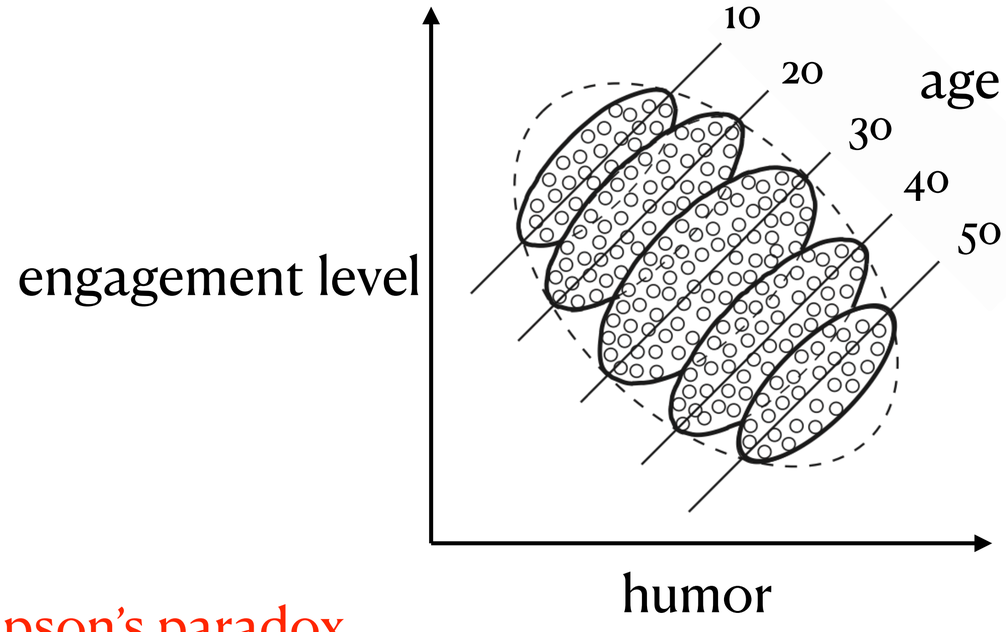


$\mathbf{y}^{(n)}(\tilde{\mathbf{a}}^{(n)})$ ?

# Challenges

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



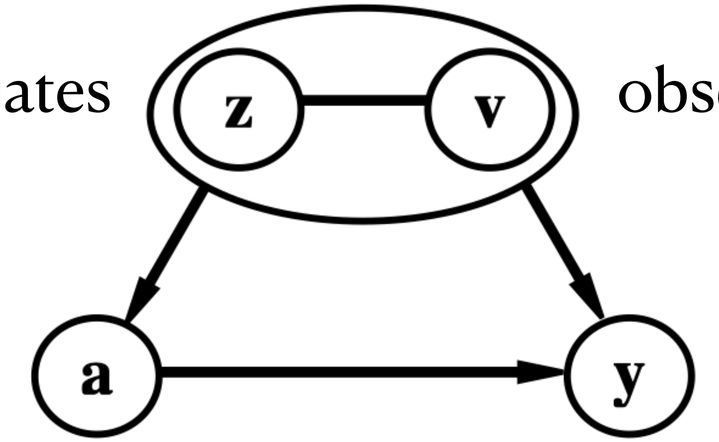




Simpson's paradox

# Problem Setup

unobserved covariates

observed covariates




observed interventions

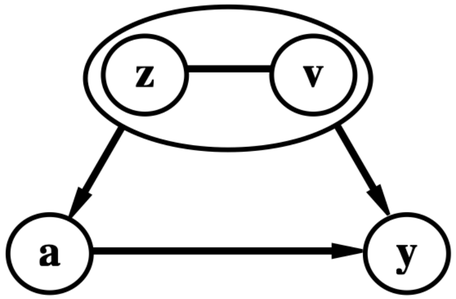
observed outcomes

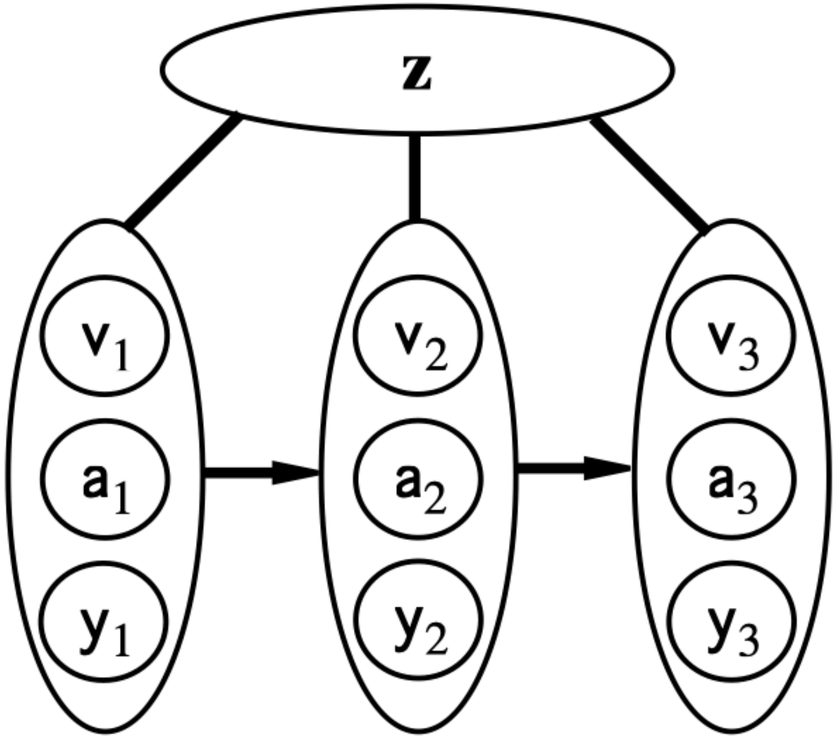
$n$ heterogenous and independent (SUTVA) users

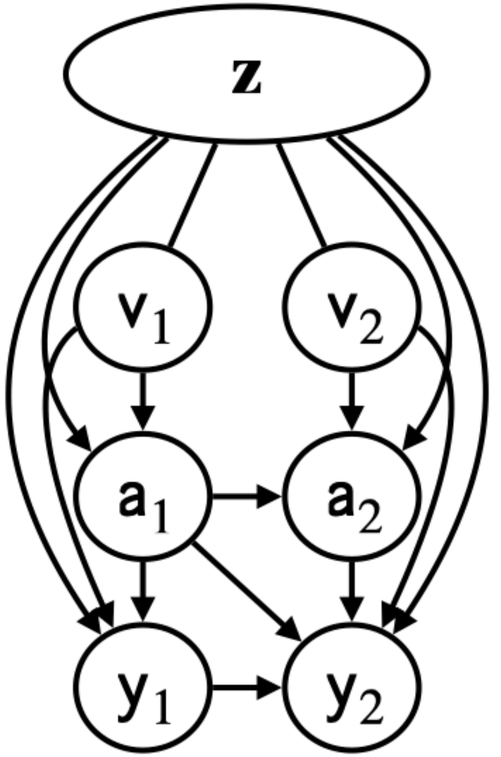
one observation per user - $\{\boldsymbol{v}^{(i)}, \boldsymbol{a}^{(i)}, \boldsymbol{y}^{(i)}\}_{i=1}^{n}$ (high-dim)

# Examples

## Sequential setting

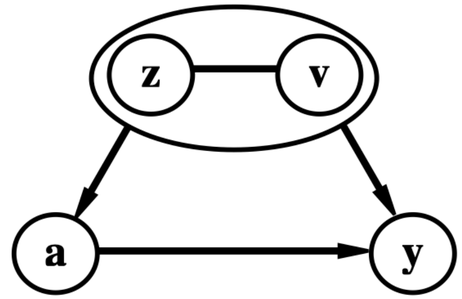


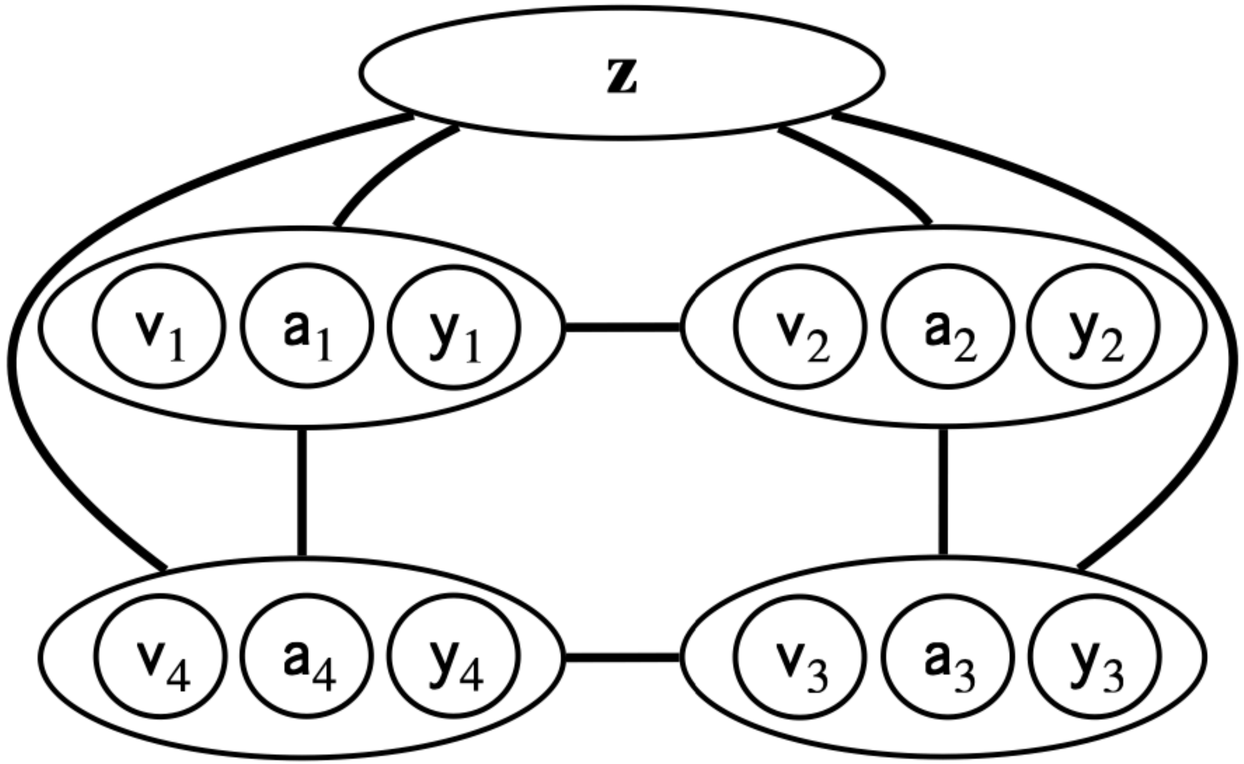



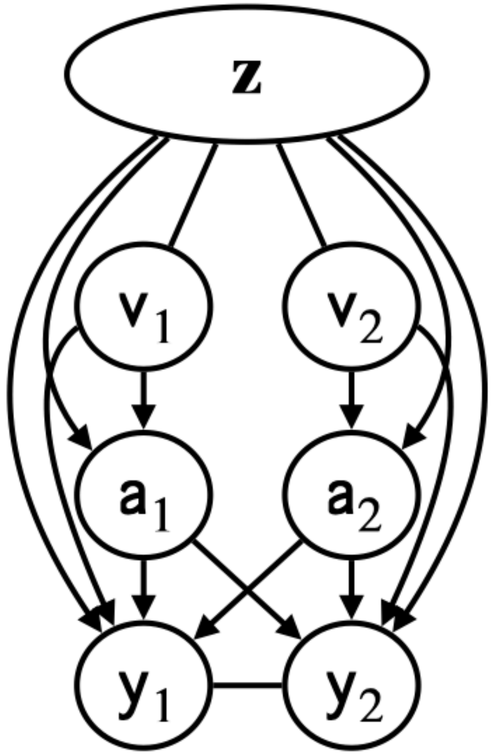



# Examples

## Network setting



# Goal: Counterfactual Questions

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

$\equiv$

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

suffices to learn $p(\mathbf{y} = \cdot \mid \mathbf{a} = \cdot, z^{(i)}, \boldsymbol{v}^{(i)})$ for all $i \in [n]$

**Heterogeneity:** each user may have different *unobserved* z

Can we learn $n$ different distributions

$p(\mathbf{y} = \cdot \mid \mathbf{a} = \cdot, z^{(i)}, \boldsymbol{v}^{(i)}) \quad i \in [n]$

with *one* sample per distribution?

# Our Approach

Model the joint distribution of $\mathbf{w} \triangleq (\mathbf{z}, \mathbf{v}, \mathbf{a}, \mathbf{y})$ as a particular exponential family

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

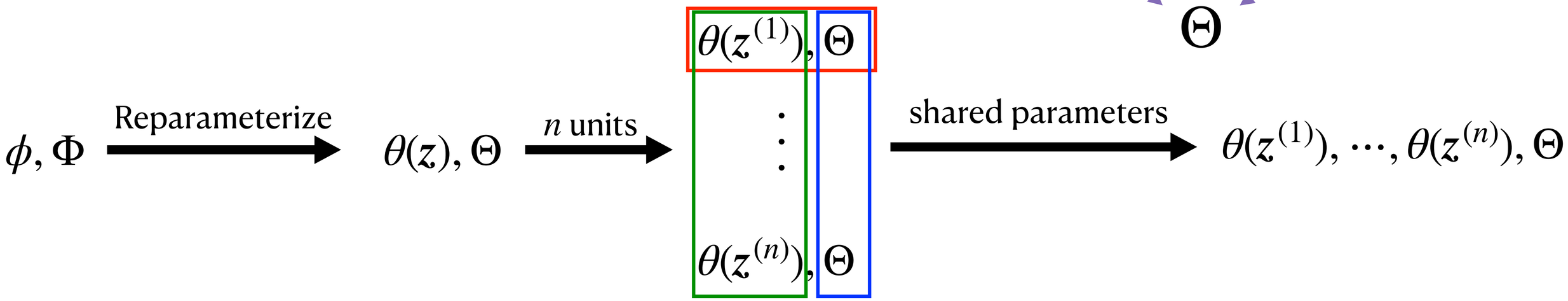
$$p(\boldsymbol{y} \,|\, \boldsymbol{a}, \boldsymbol{z} = \boldsymbol{z}^{(i)}, \boldsymbol{v} = \boldsymbol{v}^{(i)}) \propto \exp\left(\left[\, \phi_y^\top + 2\boldsymbol{z}^{(i)\top}\Phi_{z,y} + 2\boldsymbol{v}^{(i)\top}\Phi_{v,y} + 2\boldsymbol{a}^\top\Phi_{a,y} \,\right]\boldsymbol{y} + \boldsymbol{y}^\top\Phi_{y,y}\boldsymbol{y}\right)$$

different for different users

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

# Our Approach

$$p(\boldsymbol{y} \mid \boldsymbol{a}, \boldsymbol{z} = \boldsymbol{z}^{(i)}, \boldsymbol{v} = \boldsymbol{v}^{(i)}) \propto \exp\Bigg( \Big[ \underbrace{\phi_y^\top + 2\boldsymbol{z}^{(i)\top}\Phi_{z,y}}_{= \theta(\boldsymbol{z}^{(i)})} + 2\boldsymbol{v}^{(i)\top}\Phi_{v,y} + 2\boldsymbol{a}^\top\Phi_{a,y} \Big]\boldsymbol{y} + \boldsymbol{y}^\top\Phi_{y,y}\boldsymbol{y} \Bigg)$$




1. If $z^{(1)} = \cdots = z^{(n)} \rightarrow$ a single exponential family with $n$ samples **[B '15, KM '17, SSW '21, VML '22]**
2. If $n = 1 \rightarrow$ a single exponential family with one sample (assume $\Theta$ is known) **[KDDGD '21, MHBM' 21]**

# Inference Tasks

1. Parameters
   A. User-level — $\theta^\star(\boldsymbol{z}^{(i)})$ for all $i \in [n]$ → counterfactual distribution
   B. Population-level — $\Theta^\star$
2. Expected potential outcomes — $\mathbf{E}\left[\boldsymbol{y}^{(i)}(\tilde{\boldsymbol{a}}^{(i)}) \,|\, \mathbf{z} = \boldsymbol{z}^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}\right]$ → counterfactual mean

# Complexity of Parameter Space

$\Lambda_\theta$ — the set containing the true model parameters $\{\theta^\star(z^{(i)})\}_{i\in[n]}$

| | Linear combination of $k$-known vectors | $s$-sparse linear combination of $k$-known vectors |
|---|---|---|
| $M(\epsilon) = logN(\Lambda_\theta, \epsilon)$ | $O\left(k \cdot log\left(1+\frac{1}{\epsilon}\right)\right)$ | $O\left(s \log k \cdot log\left(1+\frac{1}{\epsilon}\right)\right)$ |
| $M_n(\epsilon) = nM(n\epsilon)$ | $O\left(\frac{k}{\epsilon}\right)$ | $O\left(\frac{s \log k}{\epsilon}\right)$ |

# Parameter Estimation

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

pool all $n$ samples

Loss function

estimates

$$\|\Theta^{\star} - \hat{\Theta}\|_{2,\infty} \leq \epsilon \qquad \text{when } n \geq O\left(\frac{p^2\left(p + M_n(\epsilon^2)\right)}{\epsilon^4}\right)$$

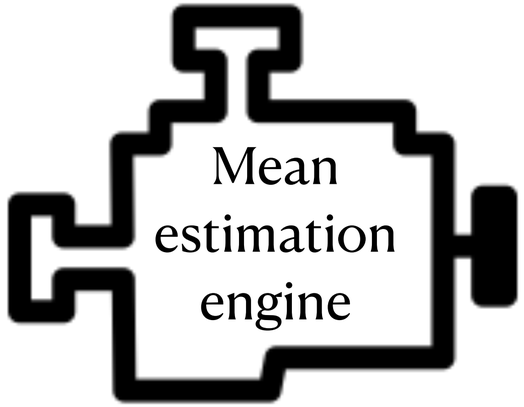
$$\text{For all } i, \text{MSE}\left(\theta^{\star}(\boldsymbol{z}^{(i)}), \hat{\theta}(\boldsymbol{z}^{(i)})\right) \leq \max\left\{\epsilon^2, \frac{M(c)}{p}\right\} \quad \text{when } n \geq O\left(\frac{p^2\left(pM(c) + M_n(\epsilon^2)\right)}{\epsilon^4}\right)$$

# Outcome Estimation

Expected potential outcomes — $\mu^{(i)} \triangleq \mathbf{E}\left[\boldsymbol{y}^{(i)}(\tilde{\boldsymbol{a}}^{(i)}) \,|\, \mathbf{z} = \boldsymbol{z}^{(i)}, \mathbf{v} = \boldsymbol{v}^{(i)}\right]$

$(\boldsymbol{v}^{(i)}, \tilde{\boldsymbol{a}}^{(i)})$

$\hat{\theta}(\boldsymbol{z}^{(i)}), \hat{\Theta}$




$\hat{\mu}^{(i)}$

For all $i$, $MSE\left(\mu^{(i)}, \hat{\mu}^{(i)}\right) \leq \epsilon^2 + \frac{M(c)}{p}$ when $n \geq O\left(\frac{p^2\left(pM(c) + M_n(\epsilon^2)\right)}{\epsilon^4}\right)$

$\mathbf{x} \triangleq (\mathbf{v}, \mathbf{a}, \mathbf{y})$

# An Application

## Denoise user-wise data

No systematically unobserved covariates $\mathbf{z}$

Noisy observed data = true data + measurement error

$\overline{\mathbf{x}}$ $\mathbf{x}$ $\Delta\mathbf{x}$

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

Assum 2: Data hasa sparse error: $\|\Delta\mathbf{x}^{(i)}\|_0 \leq s$ for $i \in \{1, \cdots, n/2\}$

Goal: Estimate the true data

# An Application

## Denoise user-wise data

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

Assum 2: Data hasa sparse error: $\|\Delta\mathbf{x}^{(i)}\|_0 \leq s$ for $i \in \{1, \cdots, n/2\}$

Our results can recover data such that

$$\text{For all } i,\ \|\Delta\mathbf{x}^{(i)}, \widehat{\ \Delta\mathbf{x}^{(i)}\ }\|^2 \leq \max\left\{\frac{\epsilon^2}{s}, \frac{s}{p}\right\} + \epsilon^2 \quad \text{when } n \geq O\left(\frac{s^2 p}{\epsilon^4}\right)$$

# Remainder Of The Talk

## The Loss Function, And Why It Works

Log likelihood as the loss function = Max. Likelihood Est.

not the right answer (computational challenge)

An alternative

a *proper* loss function, computationally efficient

ingredients of analysis

*population-level* parameter est through concentration due to many samples

*user-level* parameter est through *single* sample concentration due to

Log Sobolev + Dobrushin's criteria

# Maximum Likelihood Estimation

Consider an exponential family $p_\theta(\boldsymbol{x}) = \dfrac{\exp(\theta^\top f(\boldsymbol{x}))}{Z(\theta)}$

Given independent $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \sim p_{\theta^\star}(\boldsymbol{x})$, learn $\theta^\star$

$$\widehat{\theta}_{MLE} \triangleq \text{argmax}_\theta \text{ Log-Likehood}(\theta)$$

$$\text{Log-likelihood}(\theta) = \frac{1}{n} \sum_{i \in [n]} \theta^\top f(\boldsymbol{x}_i) - \log Z(\theta)$$

A. ✓ Consistency ?

B. ✓ Asymptotic normality ?

C. ✓ Asymptotic efficiency ?

D. ✗ Computational tractability ? How to compute $Z(\theta)$ ?

# An Alternative

Consider an exponential family $p_\theta(\boldsymbol{x}) = \dfrac{\exp(\theta^\top f(\boldsymbol{x}))}{Z(\theta)}$

Given independent $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \sim p_{\theta^\star}(\boldsymbol{x})$, learn $\theta^\star$

$$\widehat{\theta} \triangleq \operatorname{argmin}_\theta \ell(\theta)$$

$$\ell(\theta) = \frac{1}{n} \sum_{i \in [n]} \exp\big( - \theta^\top f(\boldsymbol{x}_i)\big)$$

likelihood

$$\prod_{i \in [n]} \frac{1}{Z(\theta)} \exp\big(\theta^\top f(\boldsymbol{x}_i)\big)$$

Convex constraint on $\theta$ $\longrightarrow$ convex optimization problem

avoids $Z(\theta)$

but is it any good?

# A Proper Loss Function

$$\ell(\theta) = \frac{1}{n} \sum_{i \in [n]} \exp\Big( - \theta^\top f(\boldsymbol{x}_i) \Big)$$

$$\mathbb{E}_{\theta^\star}\Big[\ell(\theta)\Big] = \mathbb{E}_{\theta^\star}\Big[\exp\Big( - \theta^\top f(\boldsymbol{x}) \Big)\Big]$$

**Theorem [SSW '21].** $\operatorname{argmin}_\theta \mathbb{E}_{\theta^\star}\Big[\ell(\theta)\Big] = \operatorname{argmin}_\theta KL(\mathcal{U} \| p_{\theta^\star - \theta})$

$\mathcal{U} \to$ uniform distribution

$\mathbb{E}_{\theta^\star}\Big[\ell(\theta)\Big]$ is minimized uniquely when $\theta = \theta^\star$

# Properties

**[SSW'21]**

$$\hat{\theta} \in \arg\min_{\theta} \ell(\theta) = \frac{1}{n} \sum_{i \in [n]} \exp\big( - \theta^\top f(\boldsymbol{x}_i) \big)$$

A. Consistency ?

B. Asymptotic normality ?

C. Asymptotic efficiency ?

D. Computational tractability ?

E. Finite sample guarantees ?

$\hat{\theta}$ is MLE w.r.t. $p_{\theta^\star - \theta}$ (not $p_\theta$)

No need to compute $Z(\theta)$ !

# Finite Sample Guarantees

**[SSW'21]**

$$\hat{\theta} \in \arg\min_{\theta} \ell(\theta) = \frac{1}{n} \sum_{i \in [n]} \exp\big( - \theta^{\top} f(\boldsymbol{x}_i) \big)$$

$$\| \theta^{\star} - \widehat{\theta} \|_2 \sim n^{-1/4}$$

Proof ingredients —

A. Concentration of gradient

B. Anti-concentration of Hessian (Restricted strong convexity)

} Hoeffding's inequality

# Back To Our Setting

## Condition on z

Recall the joint distribution of $\mathbf{w} = (\mathbf{z}, \mathbf{v}, \mathbf{a}, \mathbf{y})$

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

Letting $\mathbf{x} \triangleq (\mathbf{v}, \mathbf{a}, \mathbf{y})$, the conditional distribution of $\mathbf{x}$ given $\mathbf{z}$ can be written as

$$p_{\theta(z),\Theta}(\boldsymbol{x} \mid z) \propto \exp\left(\left[\theta(z)\right]^\top \boldsymbol{x} + \boldsymbol{x}^\top \Theta \boldsymbol{x}\right)$$

$$p(\boldsymbol{y} \mid \boldsymbol{a}, z = z^{(i)}, \boldsymbol{v} = \boldsymbol{v}^{(i)}) \propto \exp\left(\left[\phi_y + 2\boldsymbol{z}^{(i)\top}\Phi_{z,y} + 2\boldsymbol{v}^{(i)\top}\Phi_{v,y} + 2\boldsymbol{a}^\top\Phi_{a,y}\right]\boldsymbol{y} + \boldsymbol{y}^\top\Phi_{y,y}\boldsymbol{y}\right)$$

# Structure On Parameters

$$p_{\theta(z),\Theta}(x \mid z) \propto \exp\Big( \big[\theta(z)\big]^\top x + x^\top \Theta x \Big)$$

(A) Every element of $\theta^\star(z^{(1)}), \cdots, \theta^\star(z^{(n)})$, and $\Theta^\star$ are bounded

(B) Every row of $\Theta^\star$ is sparse

$\Lambda_\theta \triangleq \{\theta : \theta \text{ is consistent with (A) + low complexity}\}$ → required to provide meaningful guarantees

$\Lambda_\Theta \triangleq \{\Theta : \Theta \text{ is consistent with (A) and (B)}\}$

# Learning Population-level Parameter

$$\mathscr{L}(\underline{\Theta}) = \sum_{t\in[p]} \frac{1}{n} \sum_{i\in[n]} \exp\Big( - \big[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}\big] x_t^{(i)} \Big) \text{ where } \underline{\Theta} \triangleq \big[\theta^{(1)}, \cdots, \theta^{(n)}, \Theta\big]$$

$\Lambda_\Theta \rightarrow$ (A) bounded elements, (B) sparse rows

$\Lambda_\Theta$ places independent constraints on the rows of $\Theta$

$p$ independent optimization problems

$$\mathscr{L}_t(\underline{\Theta}_t) = \frac{1}{n} \sum_{i\in[n]} \exp\Big( - \big[\theta_t^{(i)} + 2\Theta_t^\top \boldsymbol{x}^{(i)}\big] x_t^{(i)} \Big) \text{ for all } t \in [p] \longrightarrow \widehat{\Theta}_t$$

$$\|\Theta_t^\star - \widehat{\Theta}_t\|_{2,\infty} \sim n^{-1/4}$$

A. Concentration of gradient
B. Anti-concentration of Hessian

} Hoeffding's inequality

# Learning Unit-level Parameter

$$\mathscr{L}(\theta^{(1)}, \cdots, \theta^{(n)}) = \sum_{t\in[p]} \frac{1}{n} \sum_{i\in[n]} \exp\Big( - \big[\theta_t^{(i)} + 2\,\widehat{\Theta}_t^{\top} \boldsymbol{x}^{(i)}\big] x_t^{(i)} \Big)$$

$\Lambda_\theta \to$ (A) bounded elements, (B) low complexity

$\theta^{(1)}, \cdots, \theta^{(n)} \in \Lambda_\theta^n$ places independent constraints on units, i.e., $\theta^{(i)} \in \Lambda_\theta$ for all $i \in [n]$

$n$ independent optimization problems

$$\mathscr{L}^{(i)}(\theta^{(i)}) = \sum_{t\in[p]} \exp\Big( - \big[\theta_t^{(i)} + 2\,\widehat{\Theta}_t^{\top} \boldsymbol{x}^{(i)}\big] x_t^{(i)} \Big) \text{ for all } i \in [n] \longrightarrow \widehat{\theta}^{(i)}$$

$$\|\theta^\star(\boldsymbol{z}^{(i)}) - \widehat{\theta}^{(i)}\|_2 \sim \max\{n^{-1/4}, M\}$$

A. Concentration of gradient
B. Anti-concentration of Hessian

} Logarithmic Sobolev inequality