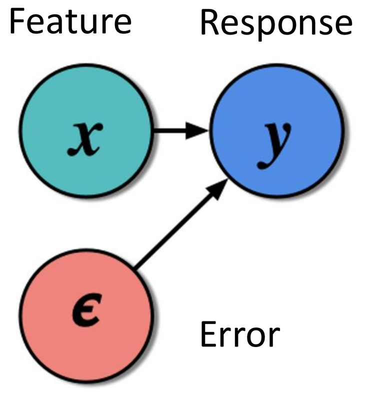
Stochastic Optimization Algorithms for Instrumental Variable Regression with Streaming Data

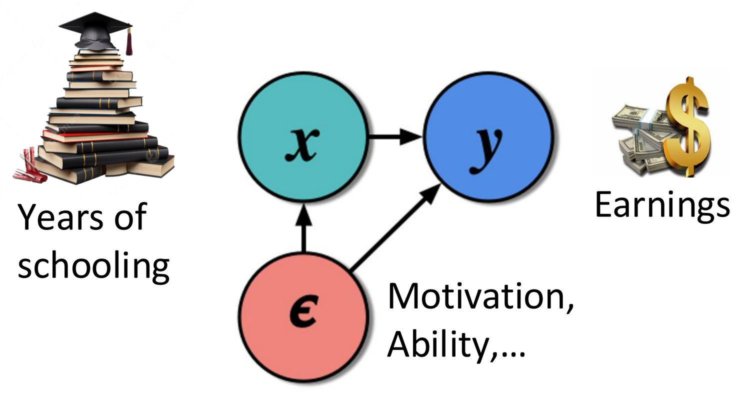
NEURAL INFORMATION PROCESSING SYSTEMS

[Xuxing Chen* (Meta), Abhishek Roy* (TAMU), Yifan Hu (EPFl, ETH Zurich), Krishna Balasubramanian (UC Davis)]

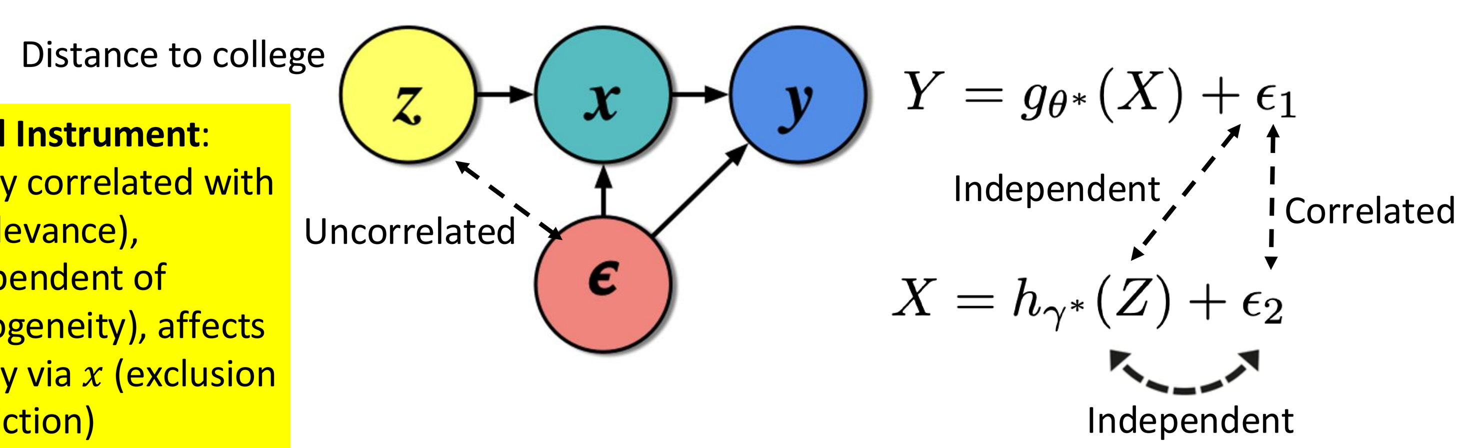
Instrumental Variable Regression



Typical linear Regression



Error in Variables, Unobserved Confounder



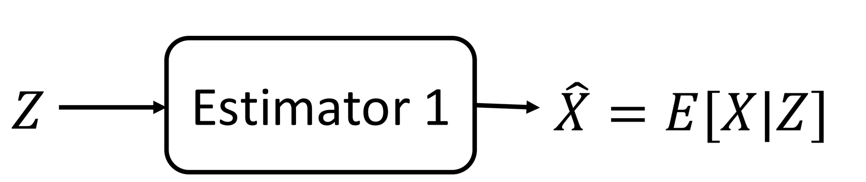
Good Instrument: Highly correlated with x (relevance), independent of ϵ(exogeneity), affects y only via x (exclusion restriction)

Instrumental Variable Regression

Estimate θ^* with streaming data?

Traditional Two-stage Method

Stage 1. Regress X on Z, obtain $\hat{X} = E[X|Z]$



Caution: Model misspecification!

Stage 2. Regress Y on $\hat{X}$ ($\hat{X}$ is uncorrelated with ϵ)

IVaR: An Optimization Viewpoint

$$\min_{\theta\in\Theta} F(\theta) = \mathbb{E}_Z \mathbb{E}_{Y|Z}[(Y - \underbrace{\mathbb{E}_{X|Z}[g_\theta(X)]}_{h_\theta(Z)})^2] \quad \text{(IVaR-Opt)}$$

Squared Loss $\Longrightarrow$ $h_{\theta^*}(Z) = \mathbb{E}[Y|Z]$

No explicit $X - Z$ model. No $X - Z$ misspecification

Challenges

- ⚠ Unknown inner expectation
- ⚠ Streaming data, can't estimate $\mathbb{E}_{X|Z}[g(X)]$
- ⚠ Biased Gradient $\nabla F(\theta, W_t) = (g(\theta_t; X_t) - Y_t)\nabla_\theta g(\theta_t; X_t)$

Our Contribution I: Two-sample Gradient Estimator

Sample: $Z_t \sim \mathcal{P}(Z)$, independent $X_t, X_t' \sim \mathcal{P}(X|Z_t)$, $Y_t \sim \mathcal{P}(Y|X_t)$

$$\nabla F(\theta_t, X_t, X_t', Y_t, Z_t) = (g(\theta_t; X_t) - Y_t)\nabla_\theta g(\theta_t; X_t') \quad \text{(Unbiased)}$$

$$\theta_{t+1} = \theta_t - \alpha_{t+1}\nabla F(\theta_t, X_t, X_t', Y_t, Z_t)$$

Theorem. ***(Squared Loss)*** *Assumptions: Identifiability, bounded moment, i.i.d data stream. Set* $\alpha_t \equiv \alpha = \frac{\log T}{\mu T}$.

$$\mathbb{E}[\|\theta_T - \theta_*\|^2] \le \frac{\|\theta_0 - \theta_*\|^2}{T} + \frac{3\|\theta_*\|^2(\sigma_1^2(d_x, d_z) + \sigma_2^2(d_x, d_z))\log T}{\mu^2 T}.$$

(General Loss) *Additional Assumptions: ℓ-Smooth F, bounded iterates. Set* $\alpha_t \equiv \alpha = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$.

$$\min_{1\le t\le T} \mathbb{E}\left[\|\nabla F(\theta_t)\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{T}}\right).$$

Takeaway: *(IVaR-Opt) is solvable with the two-sample unbiased gradient estimator, avoiding matrix inversion and explicit X-Z modeling.*

Our Contribution II: One-sample Gradient Estimator

$$Y = \theta_*^\top X + \epsilon_1 \qquad X = \gamma_*^\top Z + \epsilon_2$$

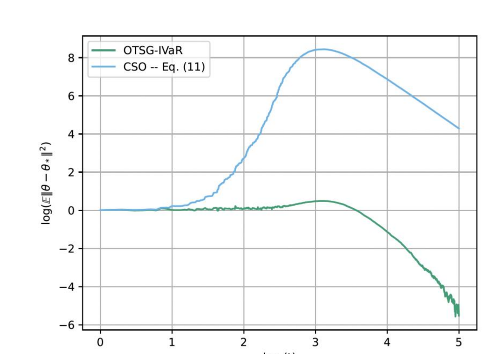
$$\gamma_{t+1} = \gamma_t - \beta_{t+1} Z_t (Z_t^\top \gamma_t - X_t^\top)$$

$$\theta_{t+1} = \theta_t - \alpha_{t+1}(\theta_t^\top X_t - Y_t)\gamma_t^\top Z_t \overset{\gamma_t \to \gamma_*}{\approx} \theta_t - \alpha_{t+1}\underbrace{(\theta_t^\top X_t - Y_t)\gamma_*^\top Z_t}_{\text{Unbiased}}$$

$$\theta_{t+1} = (I - \alpha_{t+1}\gamma_t^\top Z_t Z_t^\top \gamma_*)\theta_t - \alpha_{t+1}\gamma_t^\top Z_t(\epsilon_{2,t}^\top \theta_t - Y_t) \qquad \text{(CSO)}$$

Potentially $\preccurlyeq 0$ $\Longrightarrow$ Potential instability near bad initialization

Replace inner X_t by $\gamma_t^\top Z_t$ $\Longrightarrow$ $\theta_{t+1} = \theta_t - \alpha_{t+1}\gamma_t^\top Z_t(Z_t^\top \gamma_t \theta_t - Y_t)$ (OTSG-IVaR)



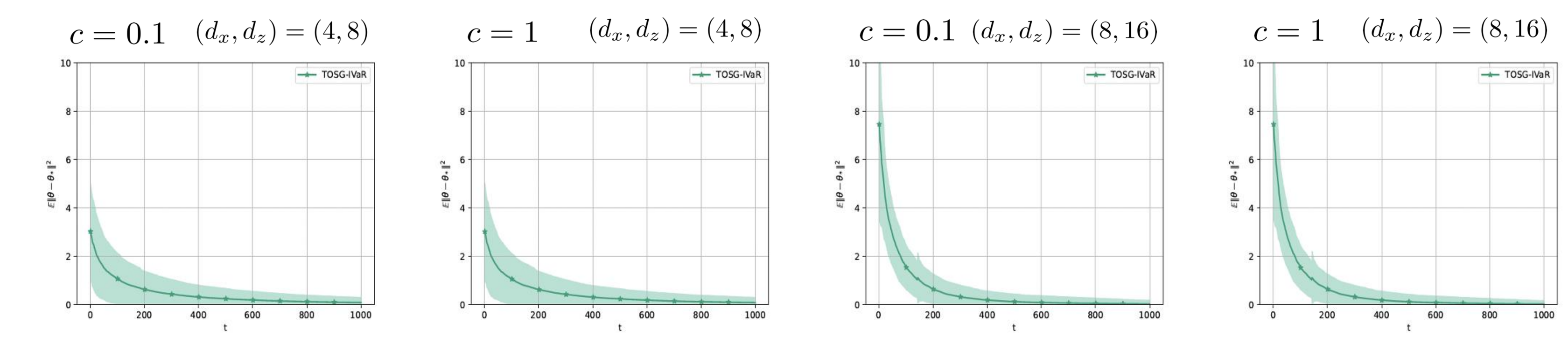
Theorem. ***(Squared Loss)*** *Assumptions: Linear models, i.i.d data stream, bounded iterates, $\Sigma_Z \succ 0$, bounded second moment. Set $\alpha_t = C_\alpha(d_z)t^{-1+\iota/2}$ and $\beta_t = C_\beta(d_z)t^{-1+\iota/2}$. Using **one sample** (X_t, Y_t, Z_t) at time t, for any $\iota > 0$, we have*

$$\mathbb{E}\left[\|\theta_t - \theta^*\|^2\right] = O\left(\frac{1}{t^{1-\iota}}\right).$$

Takeaway: *Linear IVaR is solvable with the one-sample-based gradient estimator by carefully controlling the bias, avoiding matrix inversion.*

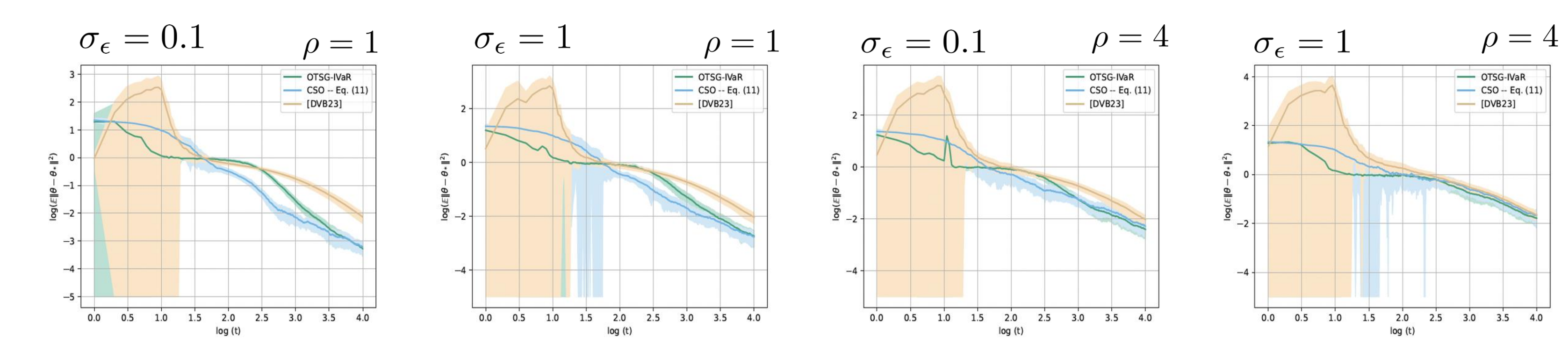
Simulation: Two Sample

$X = (\gamma_*^\top Z)^2 + c(h + \epsilon_x)$, $Y = \theta_*^\top X + c(h_1 + \epsilon_y)$ $h \sim \mathcal{N}(\mathbf{1}_{d_x}, I_{d_x})$ $Z, \epsilon_x, \epsilon_y \sim$ Standard Normal



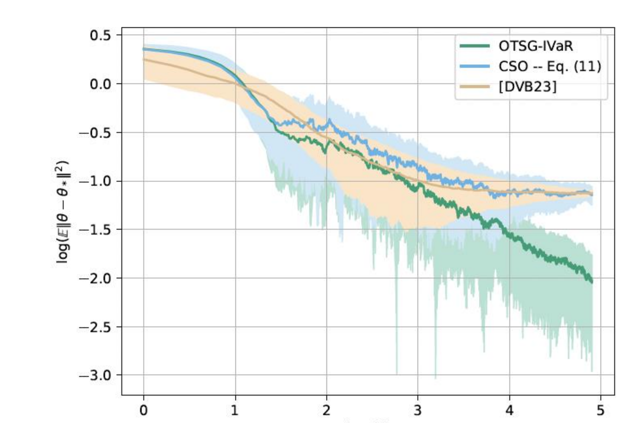
Simulation: One Sample

$Y = \theta_*^\top X + \nu$, $X = \gamma_*^\top Z + \epsilon$, $\epsilon = \sigma_\epsilon \mathcal{N}(0, I_{d_x})$, $\nu = \rho\epsilon_1 + \mathcal{N}(0, 0.25)$



Data Example I, One Sample: Children and Their Parents' Labor Supply Data in [AE96]

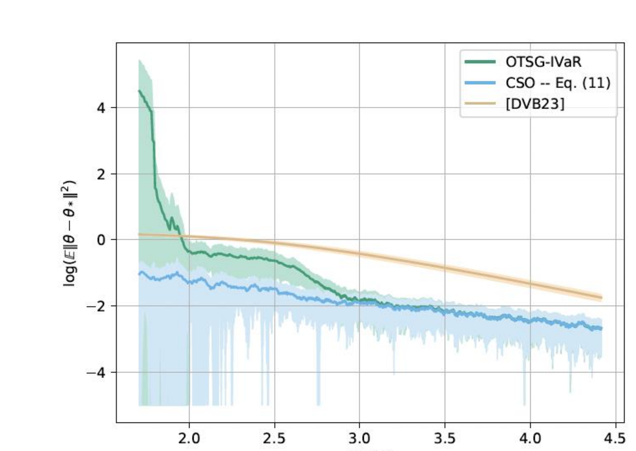
Y = number of working weeks divided by 52, $X = \mathbb{I}$(number of children is greater than 2), $Z = \mathbb{I}$(first two siblings are of same sex), $\theta_* =$ Offline estimate



OTSG-IVaR converges faster, and doesn't plateau

Data Example II, One Sample : U.S. Portland Cement Industry Data in [Rya12]

$Z = (Wage\ for\ skilled\ workers, electricity\ price, coal\ price,\ gas\ price)$, $Y = \log(shipped)$, $X = \log(price)$



OTSG-IVaR and CSO both converge faster than [DVB23]

References

[MMLR20] https://tinyurl.com/re83emkc

[DVB23] https://arxiv.org/pdf/2302.09357

[AE96] https://www.nber.org/papers/w5778

[Rya12] https://www.jstor.org/stable/41493843