February 22, 2022

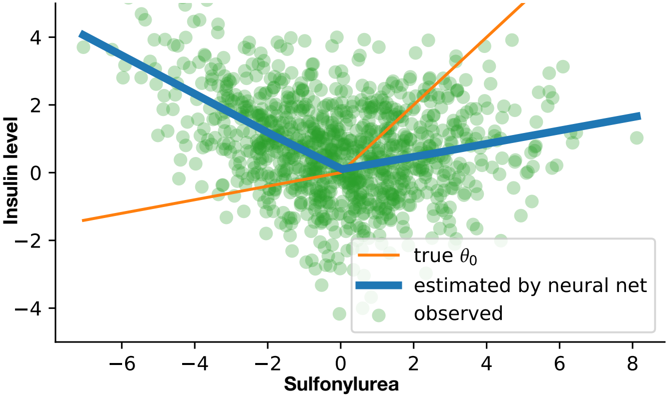
Simons Institute

The Variational Method of Moments

Nathan Kallus
Cornell University

Joint work with Andrew Bennett

Endogeneity



- $\theta_0(x) = \max(x,\, x/5)$
- $Y = \theta_0(X) - 2\epsilon + \eta, \quad \epsilon, \eta \sim \mathcal{N}(0,1)$
- $X = Z + 2\epsilon, \quad Z \sim \mathcal{N}(0,1)$

IV Model

- $Y = \theta_0(X) + \epsilon$
  - Endogeneity: $\mathbb{E}[\epsilon \mid X] \neq 0$
  - Hence, $\underbrace{\theta_0(X)}_{\text{structural/generative/causal model}} \neq \underbrace{\mathbb{E}[Y \mid X]}_{\text{predictive model}}$
- Instrument Z has
  - Exclusion: $\mathbb{E}[\epsilon \mid Z] = 0$
    - "Affects Y only via X"
  - Relevance: $\mathbb{E}[\theta(X) \mid Z] = 0 \implies \theta(X) = 0 \ \forall \theta \in \Theta - \{\theta_0\}$
    - "Affects X"
- Then θ_0 uniquely solves $\mathbb{E}[Y - \theta(X) \mid Z] = 0$ over $\theta \in \Theta$

IV is Workhorse of Empirical Research

Outcome Variable	*Endogenous Variable*	*Source of Instrumental Variable(s)*	*Reference*
		1. Natural Experiments	
Labor supply	Disability insurance replacement rates	Region and time variation in benefit rules	Gruber (2000)
Labor supply	Fertility	Sibling-Sex composition	Angrist and Evans (1998)
Education, Labor supply	Out-of-wedlock fertility	Occurrence of twin births	Bronars and Grogger (1994)
Wages	Unemployment insurance tax rate	State laws	Anderson and Meyer (2000)
Earnings	Years of schooling	Region and time variation in school construction	Duflo (2001)
Earnings	Years of schooling	Proximity to college	Card (1995)
Earnings	Years of schooling	Quarter of birth	Angrist and Krueger (1991)
Earnings	Veteran status	Cohort dummies	Imbens and van der Klaauw (1995)
Earnings	Veteran status	Draft lottery number	Angrist (1990)
Achievement test scores	Class size	Discontinuities in class size due to maximum class-size rule	Angrist and Lavy (1999)
College enrollment	Financial aid	Discontinuities in financial aid formula	van der Klaauw (1996)
Health	Heart attack surgery	Proximity to cardiac care centers	McClellan, McNeil and Newhouse (1994)
Crime	Police	Electoral cycles	Levitt (1997)
Employment and Earnings	Length of prison sentence	Randomly assigned federal judges	Kling (1999)
Birth weight	Maternal smoking	State cigarette taxes	Evans and Ringel (1999)

From Angrist & Krueger 2001

Conditional Moment Problem

- θ_0 uniquely solves the following over $\theta \in \Theta$

$$\mathbb{E}\left[\rho(O;\theta) \mid Z\right] = \mathbf{0}_m$$

- Observe $O_1, \ldots, O_n \sim O$, Z is O-measurable

Conditional Moment Problem

- θ_0 uniquely solves the following over $\theta \in \Theta$

$$\mathbb{E}\left[\rho(O;\theta) \mid Z\right] = \mathbf{0}_m$$

- Observe $O_1, \ldots, O_n \sim O$, Z is O-measurable
- Examples:
  - IV: $O = (Z, X, Y)$, $m = 1$
  - BLP model in industrial organization (Berry et al., 1995)
  - q-functions and marginal density ratios in offline RL (Liu et al., 2018, Nachum et al., 2019; Kallus & Uehara, 2019)
  - Policy learning with surrogate loss (Bennett & Kallus, 2020)
  - Proximal causal inference (Cui et al. 2020)
  - Panel data with confounders (Imbens et al., 2021)
    - Example with many θ_0's, regularization to target minimal one
  - ...

Reduction to Marginal Moment Problem

- Fix $f_j : \mathcal{Z} \to \mathbb{R}^m$, $j = 1, \ldots, k$
  - $F(z) = (f_1(z), \ldots, f_k(z)) \in \mathbb{R}^{k \times m}$
- Find $\theta_0 \in \Theta$ satisfying

$$\mathbb{E}\left[F(Z)\rho(O;\theta)\right] = \left(\mathbb{E}\left[f_j(Z)\rho(O;\theta)\right]\right)_{j=1}^{k} = \mathbf{0}_k$$

Reduction to Marginal Moment Problem

- Fix $f_j : \mathcal{Z} \to \mathbb{R}^m,\ j = 1, \ldots, k$
  - $F(z) = (f_1(z), \ldots, f_k(z)) \in \mathbb{R}^{k \times m}$
- Find $\theta_0 \in \Theta$ satisfying

$$\mathbb{E}\left[F(Z)\rho(O;\theta)\right] = \left(\mathbb{E}\left[f_j(Z)\rho(O;\theta)\right]\right)_{j=1}^{k} = \mathbf{0}_k$$

- Solve using **Optimally-Weighted Generalized Method of Moments** (OWGMM; Hansen, 1982)

$$\hat{\theta}_n \in \underset{\theta \in \Theta}{\operatorname{argmin}}\, \mathbb{E}_n[F(Z)\rho(O;\theta)]^\top \hat{\Gamma}_n^{-1}(\tilde{\theta}_n)\mathbb{E}_n[F(Z)\rho(O;\theta)],$$

where $\hat{\Gamma}_n(\tilde{\theta}_n) = \mathbb{E}_n[F(Z)\rho(O;\tilde{\theta}_n)\rho(O;\tilde{\theta}_n)^\top F(Z)^\top]$

($\mathbb{E}_n$ is the empirical average: $\mathbb{E}_n[h(O)] = \frac{1}{n}\sum_{i=1}^{n} h(O_i)$)

OWGMM

$$\hat{\theta}_n \in \underset{\theta \in \Theta}{\operatorname{argmin}}\, \mathbb{E}_n[F(Z)\rho(O;\theta)]^\top \hat{\Gamma}_n^{-1}(\tilde{\theta}_n)\mathbb{E}_n[F(Z)\rho(O;\theta)]$$

- Benefits:

OWGMM

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_n[F(Z)\rho(O;\theta)]^\top \hat{\Gamma}_n^{-1}(\tilde{\theta}_n)\mathbb{E}_n[F(Z)\rho(O;\theta)]$$

- Benefits:
  - Consistent and asymptotically normal if θ_0 uniquely solves $\mathbb{E}\left[F(Z)\rho(O;\theta)\right] = \mathbf{0}_k$ over $\theta \in \Theta$

OWGMM

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_n[F(Z)\rho(O;\theta)]^\top \hat{\Gamma}_n^{-1}(\tilde{\theta}_n)\mathbb{E}_n[F(Z)\rho(O;\theta)]$$

- Benefits:
  - Consistent and asymptotically normal if θ_0 uniquely solves $\mathbb{E}\left[F(Z)\rho(O;\theta)\right] = \mathbf{0}_k$ over $\theta \in \Theta$
  - Efficient in the model satisfying $\mathbb{E}\left[F(Z)\rho(O;\theta_0)\right] = \mathbf{0}_k$ (if $\tilde{\theta}_n \to \theta_0$)

OWGMM

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_n[F(Z)\rho(O;\theta)]^\top \hat{\Gamma}_n^{-1}(\tilde{\theta}_n)\mathbb{E}_n[F(Z)\rho(O;\theta)]$$

- Benefits:
  - Consistent and asymptotically normal if θ_0 uniquely solves $\mathbb{E}[F(Z)\rho(O;\theta)] = \mathbf{0}_k$ over $\theta \in \Theta$
  - Efficient in the model satisfying $\mathbb{E}[F(Z)\rho(O;\theta_0)] = \mathbf{0}_k$ (if $\tilde{\theta}_n \to \theta_0$)
- Limitations:

OWGMM

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_n[F(Z)\rho(O;\theta)]^\top \hat{\Gamma}_n^{-1}(\tilde{\theta}_n)\mathbb{E}_n[F(Z)\rho(O;\theta)]$$

- Benefits:
  - Consistent and asymptotically normal if θ_0 uniquely solves $\mathbb{E}\left[F(Z)\rho(O;\theta)\right] = \mathbf{0}_k$ over $\theta \in \Theta$
  - Efficient in the model satisfying $\mathbb{E}\left[F(Z)\rho(O;\theta_0)\right] = \mathbf{0}_k$ (if $\tilde{\theta}_n \to \theta_0$)
- Limitations:
  - $\mathbb{E}\left[F(Z)\rho(O;\theta_0)\right] = \mathbf{0}_k$ might not identify θ_0 (not unique)
    - *E.g.*, almost anything that isn't linear IV

OWGMM

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_n[F(Z)\rho(O;\theta)]^\top \hat{\Gamma}_n^{-1}(\tilde{\theta}_n)\mathbb{E}_n[F(Z)\rho(O;\theta)]$$

- Benefits:
  - Consistent and asymptotically normal if θ_0 uniquely solves $\mathbb{E}\left[F(Z)\rho(O;\theta)\right] = \mathbf{0}_k$ over $\theta \in \Theta$
  - Efficient in the model satisfying $\mathbb{E}\left[F(Z)\rho(O;\theta_0)\right] = \mathbf{0}_k$ (if $\tilde{\theta}_n \to \theta_0$)
- Limitations:
  - $\mathbb{E}\left[F(Z)\rho(O;\theta_0)\right] = \mathbf{0}_k$ might not identify θ_0 (not unique)
    - *E.g.*, almost anything that isn't linear IV
  - Even if identifying, not efficient in the full conditional moment model
    - *E.g.*, almost anything that isn't linear IV with linear $\mathbb{E}[X \mid Z]$

Sieve approaches

- Sieve OWGMM (Chamberlain, 1987; Ai & Chen, 2003)
  - $F = (f_1, \ldots, f_{k_n})$ first k_n elements of basis for L_2, $k_n \to \infty$
  - *E.g.*, Hermite polynomials, Fourier basis, B-splines, ...

Sieve approaches

- Sieve OWGMM (Chamberlain, 1987; Ai & Chen, 2003)
  - $F = (f_1, \ldots, f_{k_n})$ first k_n elements of basis for L_2, $k_n \to \infty$
  - *E.g.*, Hermite polynomials, Fourier basis, B-splines, ...
- Sieve-estimate the efficient instruments (Newey, 1993)
  - $F^*(Z) = (\mathbb{E}[\rho(O;\theta)\rho(O;\theta)^\top \mid Z])^{-1}\mathbb{E}[\partial_\theta \rho(O;\theta) \mid Z]$

Sieve approaches

- Sieve OWGMM (Chamberlain, 1987; Ai & Chen, 2003)
  - $F = (f_1, \ldots, f_{k_n})$ first k_n elements of basis for L_2, $k_n \to \infty$
  - *E.g.*, Hermite polynomials, Fourier basis, B-splines, ...
- Sieve-estimate the efficient instruments (Newey, 1993)
  - $F^*(Z) = (\mathbb{E}[\rho(O;\theta)\rho(O;\theta)^\top \mid Z])^{-1}\mathbb{E}[\partial_\theta \rho(O;\theta) \mid Z]$

- Theoretically efficient (under appropriate conditions)
- Unwieldy in practice, especially when θ and Z are moderately-dimensional

Minimax approaches

$$\hat{\theta}_n \in \underset{\theta \in \Theta}{\operatorname{argmin}} \sup_{f \in \mathcal{F}} \mathbb{E}_n[f(Z)^\top \rho(O;\theta)]$$

- Given a rich class of functions $\mathcal{F} \subset [\mathcal{Z} \to \mathbb{R}^m]$
  - *E.g.*, neural nets with m outputs, product of RKHSs, ...
- Try to control all marginal moments for all $f \in \mathcal{F}$
  - Not just $f_1, \ldots, f_k$
- Lewis & Syrgkanis (2018), Bennett et al. (2019), Dikkala et al. (2020), Kallus et al. (2021), Uehara et al. (2021), ...

Minimax approaches

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \sup_{f \in \mathcal{F}} \mathbb{E}_n[f(Z)^\top \rho(O; \theta)]$$

- Benefits:
  - Identification more plausible
  - No crazy sieves; much more ML-ish
  - Rates for nonparametric Θ, $\mathcal{F}$ (Dikkala et al., 2020)

Minimax approaches

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \sup_{f \in \mathcal{F}} \mathbb{E}_n[f(Z)^\top \rho(O; \theta)]$$

- Benefits:
  - Identification more plausible
  - No crazy sieves; much more ML-ish
  - Rates for nonparametric Θ, $\mathcal{F}$ (Dikkala et al., 2020)
- Limitations:
  - Efficiency?
    - Big deal because *lots* of moments in $\mathcal{F}$
  - Inference?
    - Big deal because want to do *science*!

This talk

Variational Reformulation of OWGMM

- Given $F = (f_1, \ldots, f_k)$, $f_j : \mathcal{Z} \to \mathbb{R}^m$, recall

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_n[F(Z)\rho(O;\theta)]^\top \hat{\Gamma}_n^{-1}(\tilde{\theta}_n)\mathbb{E}_n[F(Z)\rho(O;\theta)]$$

Variational Reformulation of OWGMM

- Given $F = (f_1, \ldots, f_k)$, $f_j : \mathcal{Z} \to \mathbb{R}^m$, recall

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_n[F(Z)\rho(O;\theta)]^\top \hat{\Gamma}_n^{-1}(\tilde{\theta}_n)\mathbb{E}_n[F(Z)\rho(O;\theta)]$$

Theorem

Set $\mathcal{F} = \operatorname{span}(f_1, \ldots, f_k) = \{z \mapsto \sum_{j=1}^k \beta_j f_j(z)^\top \beta : \beta \in \mathbb{R}^k\}$
OWGMM is equivalent to

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \sup_{f \in \mathcal{F}} \mathbb{E}_n[f(Z)^\top \rho(O;\theta)] - \frac{1}{4}\mathbb{E}_n[(f(Z)\rho(O;\tilde{\theta}_n))^2]$$

- Arises by Euclidean-norm duality

Variational Reformulation of OWGMM

- Given $F = (f_1, \ldots, f_k)$, $f_j : \mathcal{Z} \to \mathbb{R}^m$, recall

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}_n[F(Z)\rho(O;\theta)]^\top \hat{\Gamma}_n^{-1}(\tilde{\theta}_n)\mathbb{E}_n[F(Z)\rho(O;\theta)]$$

Theorem

Set $\mathcal{F} = \operatorname{span}(f_1, \ldots, f_k) = \{z \mapsto \sum_{j=1}^k \beta_j f_j(z)^\top \beta : \beta \in \mathbb{R}^k\}$
OWGMM is equivalent to

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \sup_{f \in \mathcal{F}} \mathbb{E}_n[f(Z)^\top \rho(O;\theta)] - \frac{1}{4}\mathbb{E}_n[(f(Z)\rho(O;\tilde{\theta}_n))^2]$$

- Arises by Euclidean-norm duality
- VMM: just switch out $\mathcal{F}$ by other function classes ...

Variational Method of Moments

$$\hat{\theta}_n \in \underset{\theta\in\Theta}{\operatorname{argmin}} \sup_{f\in\mathcal{F}_n} \mathbb{E}_n[f(Z)^\top \rho(O;\theta)] - \frac{1}{4}\mathbb{E}_n[(f(Z)\rho(O;\tilde{\theta}_n))^2] - R_n(f)$$

Variational Method of Moments

Try to make all moments zero

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \sup_{f \in \mathcal{F}_n} \mathbb{E}_n[f(Z)^\top \rho(O;\theta)] - \frac{1}{4}\mathbb{E}_n[(f(Z)\rho(O;\tilde{\theta}_n))^2] - R_n(f)$$

Variational Method of Moments

Try to make all moments zero

Regularize moments by importance

$$\hat{\theta}_n \in \underset{\theta \in \Theta}{\operatorname{argmin}} \sup_{f \in \mathcal{F}_n} \mathbb{E}_n[f(Z)^\top \rho(O;\theta)] - \frac{1}{4}\mathbb{E}_n[(f(Z)\rho(O;\tilde{\theta}_n))^2] - R_n(f)$$

Variational Method of Moments

Try to make all moments zero

Regularize moments by importance

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta\in\Theta} \sup_{f\in\mathcal{F}_n} \mathbb{E}_n[f(Z)^\top \rho(O;\theta)] - \frac{1}{4}\mathbb{E}_n[(f(Z)\rho(O;\tilde{\theta}_n))^2] - R_n(f)$$

Control complexity

Variational Method of Moments

Try to make all moments zero

Regularize moments by importance

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \Theta} \sup_{f \in \mathcal{F}_n} \mathbb{E}_n[f(Z)^\top \rho(O;\theta)] - \frac{1}{4}\mathbb{E}_n[(f(Z)\rho(O;\tilde{\theta}_n))^2] - R_n(f)$$

Can permit approx optim

Control complexity

Variational Method of Moments

Try to make all moments zero

Regularize moments by importance

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta\in\Theta} \sup_{f\in\mathcal{F}_n} \mathbb{E}_n[f(Z)^\top \rho(O;\theta)] - \frac{1}{4}\mathbb{E}_n[(f(Z)\rho(O;\tilde{\theta}_n))^2] - R_n(f)$$

Can permit approx optim

Control complexity

- k-stage VMM: initialize some $\hat{\theta}_n^{(0)}$
  - For $j = 1, \ldots, k$, set $\hat{\theta}_n^{(j)}$ to VMM with $\tilde{\theta}_n = \hat{\theta}_n^{(j-1)}$

VMM Variants

- Kernel VMM
  - Set $\mathcal{F}_n = \mathcal{H}$ to a reproducing kernel Hilbert space (RKHS)
    - *E.g.*, Gaussian kernel, product of m Sobolev spaces
  - Set $R_n(f) = \frac{\alpha_n}{4} \|f\|_{\mathcal{H}}^2$

VMM Variants

- Kernel VMM
  - Set $\mathcal{F}_n = \mathcal{H}$ to a reproducing kernel Hilbert space (RKHS)
    - *E.g.*, Gaussian kernel, product of m Sobolev spaces
  - Set $R_n(f) = \frac{\alpha_n}{4} \|f\|_{\mathcal{H}}^2$
- Neural VMM
  - Set $\mathcal{F}_n$ to a class of neural networks with a given architecture (possibly growing with n) and unknown weights
  - Kernel regularizer: set $R_n(f) = \frac{\alpha_n}{4} \inf_{h \in \mathcal{H}: h(Z_i) = f(Z_i)\ \forall i} \|h\|_{\mathcal{H}}^2$ where $\mathcal{H}$ is a given RKHS
    - $R_n(f)$ has a closed form as a quadratic in $f(Z_i)$ in terms of kernel Gram matrix

VMM Variants

- Kernel VMM
  - Set $\mathcal{F}_n = \mathcal{H}$ to a reproducing kernel Hilbert space (RKHS)
    - *E.g.*, Gaussian kernel, product of m Sobolev spaces
  - Set $R_n(f) = \frac{\alpha_n}{4} \|f\|_{\mathcal{H}}^2$
- Neural VMM
  - Set $\mathcal{F}_n$ to a class of neural networks with a given architecture (possibly growing with n) and unknown weights
  - Kernel regularizer: set $R_n(f) = \frac{\alpha_n}{4} \inf_{h \in \mathcal{H} : h(Z_i) = f(Z_i)\ \forall i} \|h\|_{\mathcal{H}}^2$ where $\mathcal{H}$ is a given RKHS
    - $R_n(f)$ has a closed form as a quadratic in $f(Z_i)$ in terms of kernel Gram matrix
  - Frobenius regularizer: set $R_n(f) = \frac{\alpha_n}{4} \sum_{k=1}^{m} \sum_{i=1}^{n} f_k^2(Z_i)$
    - Approximates Gaussian kernel regularizer w/ tiny length scale
    - Heuristic practical version of neural VMM

Implementing VMM

- It's a smooth game!
  - So, *you* tell *me* how to solve it

Implementing VMM

- It's a smooth game!
  - So, *you* tell *me* how to solve it
- Kernel VMM: inner sup has closed form as a convex quadratic in $(\rho_j(O_i;\theta))_{i=1,j=1}^{n,m}$
  - In terms of kernel Gram matrices and $\rho_j(O_i;\tilde{\theta}_n)$
  - Can directly apply usual optimization algorithms to this

Implementing VMM

- It's a smooth game!
  - So, *you* tell *me* how to solve it
- Kernel VMM: inner sup has closed form as a convex quadratic in $(\rho_j(O_i;\theta))_{i=1,j=1}^{n,m}$
  - In terms of kernel Gram matrices and $\rho_j(O_i;\tilde{\theta}_n)$
  - Can directly apply usual optimization algorithms to this
- Neural VMM: will use OAdam (Daskalakis et al., 2017) in experiments
  - Lots of developments since and lots of opportunity to potentially improve

This talk

Some Regularity (Consistency)

- $\rho(o;\theta)$ is equi-Lipschitz in θ for all o
- $\sup_{o,\theta} \|\rho(o;\theta)\| < \infty$
- $\mathcal{Z} \subset \mathbb{R}^{d_z}$ bounded
- $\int \sqrt{\log N(\Theta, \epsilon)} < \infty$
  - (Trivial for $\Theta \subset \mathbb{R}^b$ compact)
- $\tilde{\theta}_n \to_p \tilde{\theta}$
- $\|\rho_j(\cdot;\tilde{\theta}_n) - \rho_j(\cdot;\tilde{\theta})\|_\infty = O_p(n^{-p})$ for some $0 < p \leq \frac{1}{2}$
  - (Will come for free for k-stage VMM)
- $\mathbb{E}[\lambda_{\min}(\mathbb{E}[\rho(O;\theta)\rho(O;\theta)^\top \mid Z])^{-2}] < \infty$ for all $\theta \in \Theta$

Consistency

- Set $\mathcal{H}$ as *any* smooth universal kernel (*e.g.*, Gaussian)
- Set $\alpha_n = o(1)$, $\alpha_n = \omega(n^{-p})$

Theorem

Kernel VMM with $\mathcal{F}_n = \mathcal{H}$ is consistent: $\hat{\theta}_n \to_p \theta_0$

Consistency

- Set $\mathcal{H}$ as *any* smooth universal kernel (*e.g.*, Gaussian)
- Set $\alpha_n = o(1)$, $\alpha_n = \omega(n^{-p})$

Theorem

Kernel VMM with $\mathcal{F}_n = \mathcal{H}$ is consistent: $\hat{\theta}_n \to_p \theta_0$

Corollary

Same for neural VMM with fully connected net with width and depth at least a certain amount (in paper) and with kernel regularizer given by $\mathcal{H}$

More Regularity (Asymptotic Normality)

- Suppose $\Theta \subset \mathbb{R}^b$ compact
  - (Covering assumption holds trivially)
- $\sup_{o,\theta} \|\frac{\partial}{\partial \theta_i} \rho(o;\theta)\| < \infty$
- $\sup_{o,\theta} \|\frac{\partial^2}{\partial \theta_i \partial \theta_j} \rho(o;\theta)\| < \infty$
- $\frac{\partial}{\partial \theta_i} \rho(o;\theta)$, $\frac{\partial^2}{\partial \theta_i \partial \theta_j} \rho(o;\theta)$ equi-Lipschitz
- $\{\mathbb{E}[\frac{\partial}{\partial \theta_i} \rho(O;\theta_0) \mid Z] : i = 1, \ldots, b\}$ are b linearly independent functions $\mathcal{Z} \to \mathbb{R}^m$

Asymptotic Normality

- Set $\mathcal{H}$ as any smooth universal kernel (*e.g.*, Gaussian)
- Set $\alpha_n = o(1)$, $\alpha_n = \omega(n^{-p})$

Theorem

Kernel VMM with $\mathcal{F}_n = \mathcal{H}$ is asymptotically linear ($\hat{\theta}_n = \mathbb{E}_n[\psi(O)] + o_p(n^{-1/2})$) and asymptotically normal:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}(0, V_{\tilde{\theta}})$$

Asymptotic Normality

- Set $\mathcal{H}$ as any smooth universal kernel (*e.g.*, Gaussian)
- Set $\alpha_n = o(1)$, $\alpha_n = \omega(n^{-p})$

Theorem

Kernel VMM with $\mathcal{F}_n = \mathcal{H}$ is asymptotically linear ($\hat{\theta}_n = \mathbb{E}_n[\psi(O)] + o_p(n^{-1/2})$) and asymptotically normal:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}(0, V_{\tilde{\theta}})$$

Corollary

Same for neural VMM with fully connected net with width and depth at least a certain amount (in paper) and with kernel regularizer given by $\mathcal{H}$

Semiparametric Efficiency

Theorem

V_{θ_0} (i.e., the asymptotic covariance of VMM when $\tilde{\theta}_n \to_p \theta_0$) is the semiparametric efficiency bound for θ_0 in the model consisting of all distributions satisfying $\mathbb{E}[\rho(O;\theta_0) \mid Z] = 0$

Semiparametric Efficiency

Theorem

V_{θ_0} (i.e., the asymptotic covariance of VMM when $\tilde{\theta}_n \to_p \theta_0$) is the semiparametric efficiency bound for θ_0 in the model consisting of all distributions satisfying $\mathbb{E}[\rho(O;\theta_0) \mid Z] = 0$

Corollary

k-stage kernel/neural VMM ($k \geq 2$) using a smooth universal kernel and $\alpha_n = o(1)$, $\alpha_n = \omega(1/\sqrt{n})$ is semiparametrically efficient in the conditional moment problem

In particular: minimum asymptotic MSE for $\beta^\top \theta_0$ for any β (either among regular estimators or locally minimax among all estimators)

This talk

Variational Reformulation of the Efficiency Bound

Theorem

Let V_{θ_0} be the efficiency bound. Then

$$\beta^\top V_{\theta_0}\beta = \sup_{\gamma\in\mathbb{R}^b}\inf_{f\in\mathcal{F}} \gamma^\top\beta - \frac{1}{4}\mathbb{E}[f(Z)^\top \nabla_\theta \rho(X;\theta_0)\gamma] + \frac{1}{16}\mathbb{E}[(f(Z)^\top\rho(X;\theta_0))^2]$$

- We estimate this variance using VMM-style minimax

$$\hat{v}_n^2(\beta) = \sup_{\gamma\in\mathbb{R}^b}\inf_{f\in\mathcal{H}} \gamma^\top\beta - \frac{1}{4}\mathbb{E}_n[f(Z)^\top \nabla_\theta \rho(X;\hat{\theta}_n)\gamma] + \frac{1}{16}\mathbb{E}_n[(f(Z)^\top\rho(X;\hat{\theta}_n))^2] - R_n(f)$$

Kernel VMM Inference

- Set $\mathcal{H}$ any smooth universal kernel (*e.g.*, Gaussian)
- Set $\alpha_n = o(1)$, $\alpha_n = \omega(n^{-p})$
- Set $\hat{\theta}_n$ to k-stage kernel/neural VMM ($k \geq 2$)

Theorem

Kernel VMM standard error estimate with $\mathcal{F}_n = \mathcal{H}$ has

$$\hat{v}_n^2(\beta) \to_p \beta^\top V_{\theta_0} \beta$$

Hence: $\mathbb{P}(\psi(\theta_0) \in [\psi(\hat{\theta}_n) \pm 1.96 \hat{v}_n(\nabla \psi(\hat{\theta}_n))]) \to 0.95$

- $\hat{v}_n(\beta)$ has a closed form in terms of kernel Gram matrices

This talk

MSE in Simple IV scenario

Method		n = 200	500	1,000	2,000	5,000	10,000
K-VMM $\alpha_n =$	0	> 100	8.8 ± 42.7	> 100	$.67 \pm 1.2$	$.23 \pm .29$	$.14 \pm .16$
	10^{-8}	5.1 ± 7.0	2.8 ± 3.0	2.6 ± 5.3	3.2 ± 16.5	$.25 \pm .32$	$.17 \pm .23$
	10^{-6}	5.5 ± 7.0	2.5 ± 2.7	1.7 ± 3.0	$.78 \pm 1.3$	$.24 \pm .33$	$.14 \pm .16$
	10^{-4}	5.5 ± 7.6	2.5 ± 3.2	1.8 ± 2.9	$.72 \pm 1.3$	$.25 \pm .32$	$.14 \pm .16$
	10^{-2}	6.0 ± 8.3	2.7 ± 3.1	1.7 ± 2.4	$.72 \pm 1.2$	$.26 \pm .34$	$.14 \pm .17$
	1	11 ± 21	4.1 ± 6.6	2.1 ± 2.8	$.75 \pm 1.1$	$.34 \pm .41$	$.16 \pm .21$
N-VMM $\lambda_n =$	0	2.5 ± 2.0	1.6 ± 1.9	$.93 \pm 1.2$	$.42 \pm .65$	$.16 \pm .21$	$.10 \pm .14$
	10^{-4}	2.8 ± 2.7	1.8 ± 2.0	$.81 \pm 1.1$	$.39 \pm .62$	$.18 \pm .25$	$.11 \pm .14$
	10^{-2}	2.2 ± 1.9	2.1 ± 2.6	$.74 \pm .99$	$.42 \pm .66$	$.17 \pm .23$	$.10 \pm .12$
	1	2.1 ± 2.0	2.1 ± 2.1	$.94 \pm 1.2$	$.39 \pm .65$	$.18 \pm .26$	$.11 \pm .12$
Sieve	Id	4.2 ± 6.5	2.5 ± 3.6	1.8 ± 3.0	$.68 \pm 1.0$	$.24 \pm .31$	$.15 \pm .19$
	Hom	4.2 ± 6.5	2.5 ± 3.6	1.8 ± 3.0	$.68 \pm 1.0$	$.24 \pm .32$	$.15 \pm .19$
	Het	4.3 ± 5.7	2.4 ± 3.3	1.7 ± 2.6	$.66 \pm 1.0$	$.24 \pm .31$	$.15 \pm .18$
MMR		17 ± 28	5.6 ± 9.2	2.8 ± 3.7	$.83 \pm 1.1$	$.37 \pm .45$	$.17 \pm .23$
Naïve		6.2 ± 1.3	$6.0 \pm .71$	$5.8 \pm .45$	$5.8 \pm .47$	$5.8 \pm .25$	$5.8 \pm .20$

MSE in Complex IV scenario

Method		n 200	500	1,000	2,000	5,000	10,000
K-VMM $\alpha_n =$	0	> 100	3.8 ± 5.5	> 100	$.63 \pm 1.4$	$.24 \pm .29$	$.09 \pm .18$
	10^{-8}	> 100	> 100	1.3 ± 2.2	$.63 \pm 2.0$	$.21 \pm .23$	$.06 \pm .05$
	10^{-6}	8.7 ± 22.9	2.0 ± 2.6	$.78 \pm .98$	$.35 \pm .50$	$.22 \pm .27$	$.06 \pm .05$
	10^{-4}	9.9 ± 27.6	1.9 ± 2.2	$.79 \pm .96$	$.35 \pm .45$	$.21 \pm .26$	$.05 \pm .05$
	10^{-2}	9.1 ± 19.7	2.6 ± 3.6	1.1 ± 1.3	$.40 \pm .49$	$.21 \pm .23$	$.06 \pm .06$
	1	10.1 ± 15.5	5.2 ± 7.0	3.5 ± 5.8	2.5 ± 4.7	1.6 ± 1.5	1.4 ± 1.5
N-VMM $\lambda_n =$	0	9.3 ± 3.7	5.3 ± 2.8	2.8 ± 1.6	1.9 ± 1.3	$1.2 \pm .84$	$.68 \pm .64$
	10^{-4}	8.2 ± 4.0	5.4 ± 2.5	2.9 ± 1.7	1.7 ± 1.3	$1.1 \pm .80$	$.71 \pm .68$
	10^{-2}	8.8 ± 4.1	5.6 ± 2.5	2.8 ± 2.0	1.8 ± 1.3	$1.1 \pm .83$	$.72 \pm .65$
	1	7.3 ± 2.7	4.9 ± 2.1	2.7 ± 1.9	2.0 ± 1.3	$1.1 \pm .84$	$.67 \pm .68$
Sieve	Id	> 100	> 100	> 100	> 100	> 100	> 100
	Hom	> 100	> 100	> 100	> 100	> 100	> 100
	Het	> 100	> 100	> 100	> 100	> 100	> 100
MMR		10.3 ± 1.9	10.2 ± 1.2	9.7 ± 1.2	$9.8 \pm .85$	$9.7 \pm .70$	$9.6 \pm .60$
Naïve		9.1 ± 6.7	8.8 ± 5.1	7.6 ± 3.0	7.9 ± 2.4	7.7 ± 1.2	$7.4 \pm .89$

L_2 error in Complex IV scenario

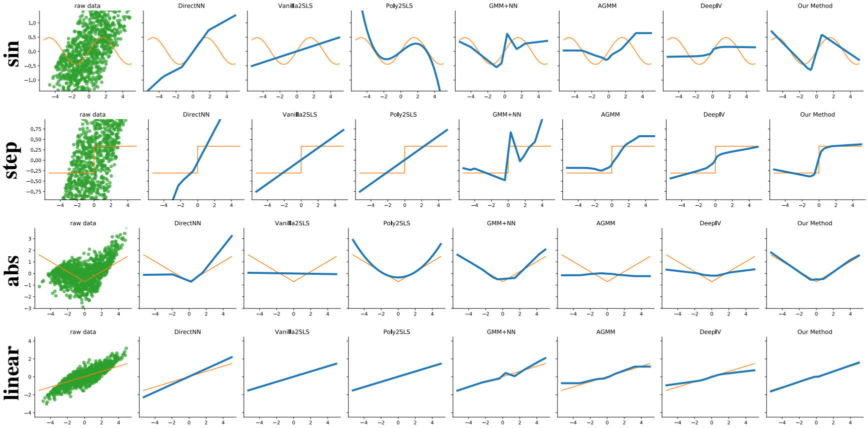
Method		n = 200	500	1,000	2,000	5,000	10,000
K-VMM $\alpha_n =$	0	> 100	$.92 \pm 1.8$	2.1 ± 13.1	$.16 \pm .34$	$.06 \pm .05$	$.03 \pm .07$
	10^{-8}	14.0 ± 59.8	> 100	$.36 \pm .69$	$.15 \pm .31$	$.05 \pm .04$	$.02 \pm .01$
	10^{-6}	1.4 ± 1.3	$.44 \pm .37$	$.19 \pm .14$	$.09 \pm .07$	$.05 \pm .04$	$.02 \pm .01$
	10^{-4}	1.4 ± 1.4	$.40 \pm .33$	$.18 \pm .13$	$.09 \pm .07$	$.05 \pm .04$	$.02 \pm .01$
	10^{-2}	1.5 ± 1.5	$.49 \pm .47$	$.21 \pm .17$	$.09 \pm .07$	$.05 \pm .03$	$.02 \pm .01$
	1	1.7 ± 1.6	$.87 \pm .79$	$.52 \pm .64$	$.35 \pm .49$	$.22 \pm .18$	$.19 \pm .19$
N-VMM $\lambda_n =$	0	5.2 ± 2.7	$1.5 \pm .74$	$.55 \pm .29$	$.32 \pm .20$	$.16 \pm .10$	$.09 \pm .07$
	10^{-4}	5.0 ± 3.0	$1.5 \pm .73$	$.58 \pm .32$	$.30 \pm .18$	$.15 \pm .09$	$.09 \pm .08$
	10^{-2}	4.8 ± 2.7	$1.5 \pm .71$	$.55 \pm .33$	$.31 \pm .19$	$.15 \pm .09$	$.09 \pm .08$
	1	3.7 ± 1.8	$1.4 \pm .58$	$.54 \pm .29$	$.32 \pm .18$	$.15 \pm .10$	$.09 \pm .08$
Sieve	Id	4.4 ± 2.9	4.4 ± 4.0	3.3 ± 3.8	2.7 ± 3.1	2.5 ± 2.9	3.7 ± 4.0
	Hom	4.3 ± 3.1	3.4 ± 5.8	3.3 ± 4.9	3.7 ± 3.9	3.6 ± 3.3	3.2 ± 3.1
	Het	4.8 ± 3.4	3.5 ± 4.0	3.4 ± 3.7	2.4 ± 2.9	3.2 ± 3.1	2.7 ± 3.3
MMR		$2.1 \pm .81$	$1.7 \pm .44$	$1.5 \pm .29$	$1.4 \pm .31$	$1.3 \pm .24$	$1.3 \pm .17$
Naïve		5.9 ± 1.3	$5.7 \pm .67$	$5.5 \pm .63$	$5.6 \pm .53$	$5.6 \pm .28$	$5.5 \pm .22$

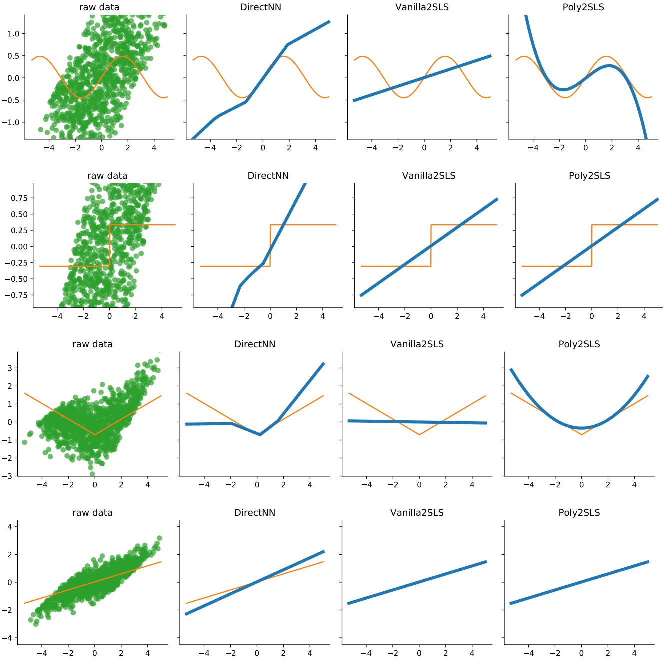
Coverage for 95% CIs

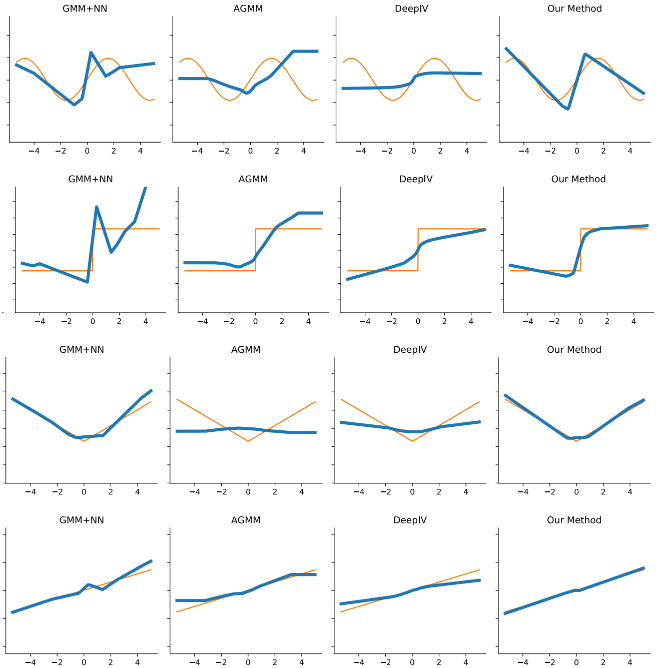
n	Method		Simple IV	Complex IV
200	Kernel	$\alpha_n = 0$	83.0	84.5
		$\alpha_n = 10^{-8}$	83.0	83.5
		$\alpha_n = 10^{-6}$	83.0	87.5
		$\alpha_n = 10^{-4}$	84.5	91.5
		$\alpha_n = 10^{-2}$	86.5	95.0
		$\alpha_n = 1$	91.0	100
	Neural	$\lambda_n = 0$	82.0	70.5
		$\lambda_n = 10^{-4}$	81.5	71.5
		$\lambda_n = 10^{-2}$	83.5	69.5
		$\lambda_n = 1$	82.5	70.0
2000	Kernel	$\alpha_n = 0$	91.5	95.5
		$\alpha_n = 10^{-8}$	92.0	95.5
		$\alpha_n = 10^{-6}$	92.5	95.5
		$\alpha_n = 10^{-4}$	92.5	96.0
		$\alpha_n = 10^{-2}$	95.0	97.5
		$\alpha_n = 1$	100.0	100
	Neural	$\lambda_n = 0$	90.0	95.5
		$\lambda_n = 10^{-4}$	90.5	95.5
		$\lambda_n = 10^{-2}$	90.0	95.5
		$\lambda_n = 1$	90.0	95.5

Beyond efficiency

- We proved VMM consistent for general θ
- But efficiency and inference only made sense for *finite-dim* θ
  - What about general nonparametric θ?
- Dikkala et al. (2020) provide nonparametric finite-sample guarantees for unweighted minimax method 👍
  - But we know plain minimax not efficient – need weighting
  - At the same time, efficiency is not a story about rates, but about leading constants on first-order terms
  - Hard to characterize the effect of optimal weighting in terms of finite-sample guarantees?
    - TBD
- But does seem to help in practice


sin
step
abs
linear
raw data
DirectNN
Vanilla2SLS
Poly2SLS
GMM+NN
AGMM
DeepIV
Our Method


raw data
DirectNN
Vanilla2SLS
Poly2SLS
raw data
DirectNN
Vanilla2SLS
Poly2SLS
raw data
DirectNN
Vanilla2SLS
Poly2SLS
raw data
DirectNN
Vanilla2SLS
Poly2SLS


GMM+NN
AGMM
DeepIV
Our Method
−4 −2 0 2 4
GMM+NN
AGMM
DeepIV
Our Method
−4 −2 0 2 4
GMM+NN
AGMM
DeepIV
Our Method
−4 −2 0 2 4
GMM+NN
AGMM
DeepIV
Our Method
−4 −2 0 2 4

This talk

Policy Learning

- Covariates X, potential losses $Y^*(+1), Y^*(-1)$
  - For $g : \mathcal{X} \to \mathbb{R}$ define

$$J(g) = \mathbb{E}[\text{sign}(g(X))(Y^*(+1) - Y^*(-1))]$$

  - Equal to (twice) the value of the policy $\text{sign}(g(X))$ minus the value of the completely randomized policy (± 1 equiprobably)

Policy Learning

- Covariates X, potential losses $Y^*(+1), Y^*(-1)$
  - For $g : \mathcal{X} \to \mathbb{R}$ define

$$J(g) = \mathbb{E}[\mathrm{sign}(g(X))(Y^*(+1) - Y^*(-1))]$$

  - Equal to (twice) the value of the policy $\mathrm{sign}(g(X))$ minus the value of the completely randomized policy (± 1 equiprobably)
- Observe $O = (X, A, Y)$ where $Y = Y^*(A)$, $A \perp\!\!\!\perp Y^*(\pm 1) \mid X$

$$J(g) = \mathbb{E}[\psi(O)\,\mathrm{sign}(g(X))]$$

$$\psi(O) = \mu(X, +1) - \mu(X, -1) + \frac{Y - \mu(X,A)}{\frac{1}{2}(A-1) + e(X)},$$
$$\mu(X, A) = \mathbb{E}[Y \mid X, A],\ e(X) = \mathbb{P}\left(A = 1 \mid X\right)$$

  - $\mathbb{E}_n[\psi(O)\,\mathrm{sign}(g(X))]$ semiparametrically efficient for $J(g)$

Reduction to Cost-Sensitive Classification

- $J(g) = \mathbb{E}[\psi(O)\,\mathrm{sign}(g(X))] = \mathbb{E}[W \ell_{\text{0-1}}\,(g(X), S)]$
  - $W = |\psi(O)|$, $S = \mathrm{sign}(\psi(O))$, $\ell_{\text{0-1}}(v,s) = \mathrm{sign}(v)s$
- For a *classification calibrated loss* ℓ (Bartlett et al., 2006):

$$g \in \operatorname{argmin} \mathbb{E}[W \ell_{\text{0-1}}(g(X), S)] \iff g \in \operatorname{argmin} \mathbb{E}[W \ell(g(X), S)]$$

Reduction to Cost-Sensitive Classification

- $J(g) = \mathbb{E}[\psi(O)\,\mathrm{sign}(g(X))] = \mathbb{E}[W\ell_{\text{0-1}}\,(g(X), S)]$
  - $W = |\psi(O)|$, $S = \mathrm{sign}(\psi(O))$, $\ell_{\text{0-1}}(v, s) = \mathrm{sign}(v)s$
- For a *classification calibrated loss* ℓ (Bartlett et al., 2006):

$$g \in \operatorname{argmin} \mathbb{E}[W\ell_{\text{0-1}}(g(X), S)] \iff g \in \operatorname{argmin} \mathbb{E}[W\ell(g(X), S)]$$

  - May restrict $g \in \mathcal{G}$ if $\mathcal{G} \cap \operatorname{argmin} \mathbb{E}[W\ell_{\text{0-1}}(g(X), S)] \neq 0$

Reduction to Cost-Sensitive Classification

- $J(g) = \mathbb{E}[\psi(O)\,\mathrm{sign}(g(X))] = \mathbb{E}[W\ell_{\text{0-1}}\,(g(X), S)]$
  - $W = |\psi(O)|$, $S = \mathrm{sign}(\psi(O))$, $\ell_{\text{0-1}}(v, s) = \mathrm{sign}(v)s$
- For a *classification calibrated loss* ℓ (Bartlett et al., 2006):

$$g \in \operatorname{argmin} \mathbb{E}[W\ell_{\text{0-1}}(g(X), S)] \iff g \in \operatorname{argmin} \mathbb{E}[W\ell(g(X), S)]$$

  - May restrict $g \in \mathcal{G}$ if $\mathcal{G} \cap \operatorname{argmin} \mathbb{E}[W\ell_{\text{0-1}}(g(X), S)] \neq 0$
- Suggests to use surrogate-loss classification

$$\hat{g}_n \in \operatorname*{argmin}_{g \in \mathcal{G}} \mathbb{E}_n[W\ell(g(X), S)]$$

  - *E.g.*, hinge (Zhou & Kosorok, '17), logistic (Jiang et al., '19)
  - For logistic can even do M-estimation inference

A Conditional Moment Problem

- $g \in \operatorname{argmin} \mathbb{E}[W\ell(g(X), S)] \iff \mathbb{E}[W\ell'(g(X), S) \mid X] = 0$

A Conditional Moment Problem

- $g \in \operatorname{argmin} \mathbb{E}[W\ell(g(X), S)] \iff \mathbb{E}[W\ell'(g(X), S) \mid X] = 0$
  - Consider $\mathcal{G} = \{g_\theta(x) = \theta^\top x : \theta \in \mathbb{R}^d\}$

A Conditional Moment Problem

- $g \in \operatorname{argmin} \mathbb{E}[W\ell(g(X), S)] \iff \mathbb{E}[W\ell'(g(X), S) \mid X] = 0$
  - Consider $\mathcal{G} = \{g_\theta(x) = \theta^\top x : \theta \in \mathbb{R}^d\}$
  - Classic MLE theory: linear logistic regression *is* efficient in the model on (X, S) satisfying $\mathbb{E}[\ell'(g(X), S) \mid X] = 0$

A Conditional Moment Problem

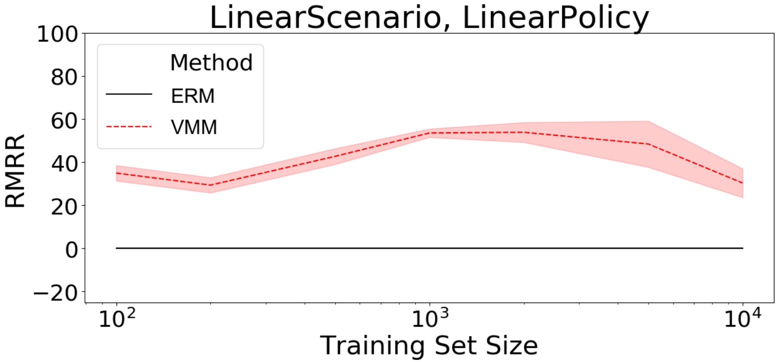
- $g \in \operatorname{argmin} \mathbb{E}[W\ell(g(X), S)] \iff \mathbb{E}[W\ell'(g(X), S) \mid X] = 0$
  - Consider $\mathcal{G} = \{g_\theta(x) = \theta^\top x : \theta \in \mathbb{R}^d\}$
  - Classic MLE theory: linear logistic regression *is* efficient in the model on (X, S) satisfying $\mathbb{E}[\ell'(g(X), S) \mid X] = 0$
  - Surprisingly, weighted logistic regression $\hat{\theta}_n \in \operatorname{argmin}_{g \in \mathcal{G}} \mathbb{E}_n[W\ell(g_\theta(X), S)]$ is *not* efficient for θ_0 in the above policy learning setting

A Conditional Moment Problem

- $g \in \operatorname{argmin} \mathbb{E}[W\ell(g(X), S)] \iff \mathbb{E}[W\ell'(g(X), S) \mid X] = 0$
  - Consider $\mathcal{G} = \{g_\theta(x) = \theta^\top x : \theta \in \mathbb{R}^d\}$
  - Classic MLE theory: linear logistic regression *is* efficient in the model on (X, S) satisfying $\mathbb{E}[\ell'(g(X), S) \mid X] = 0$
  - Surprisingly, weighted logistic regression $\hat{\theta}_n \in \operatorname{argmin}_{g \in \mathcal{G}} \mathbb{E}_n[W\ell(g_\theta(X), S)]$ is *not* efficient for θ_0 in the above policy learning setting
- Can use VMM to get efficient learner
  - Efficiency has optimal regret implications

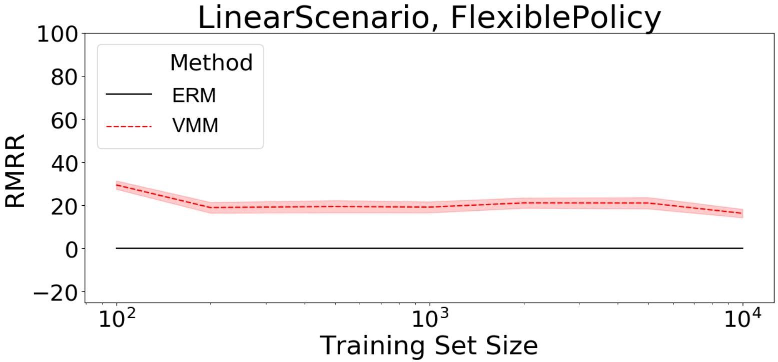
Numerics

- $\text{RMRR} = \left(1 - \frac{\mathbb{E}[J(\hat{g}^{\text{VMM}})] - \inf_g J(g)}{\mathbb{E}[J(\hat{g}^{\text{ERM}})] - \inf_g J(g)}\right) \times 100\%$



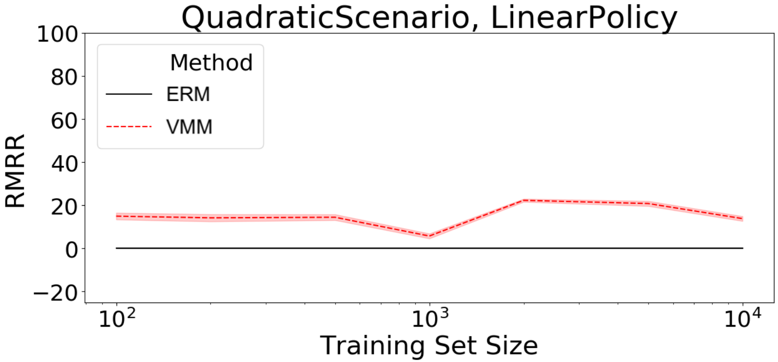
Numerics

- $$\text{RMRR} = \left(1 - \frac{\mathbb{E}[J(\hat{g}^{\text{VMM}})] - \inf_g J(g)}{\mathbb{E}[J(\hat{g}^{\text{ERM}})] - \inf_g J(g)}\right) \times 100\%$$



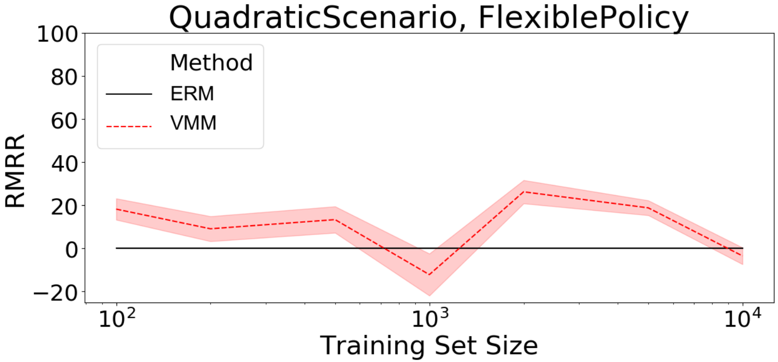
Numerics

- $\text{RMRR} = \left(1 - \frac{\mathbb{E}[J(\hat{g}^{\text{VMM}})] - \inf_g J(g)}{\mathbb{E}[J(\hat{g}^{\text{ERM}})] - \inf_g J(g)}\right) \times 100\%$



Numerics

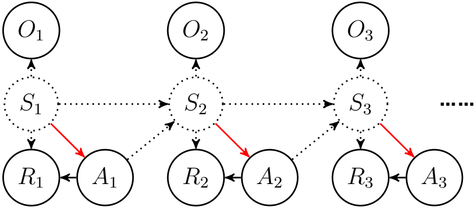
- $$\text{RMRR} = \left(1 - \frac{\mathbb{E}[J(\hat{g}^{\mathsf{VMM}})] - \inf_g J(g)}{\mathbb{E}[J(\hat{g}^{\mathsf{ERM}})] - \inf_g J(g)}\right) \times 100\%$$



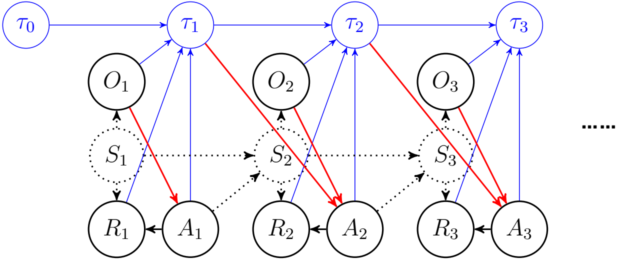
This talk

Model

- POMDP confounded observation model



- Evaluation task

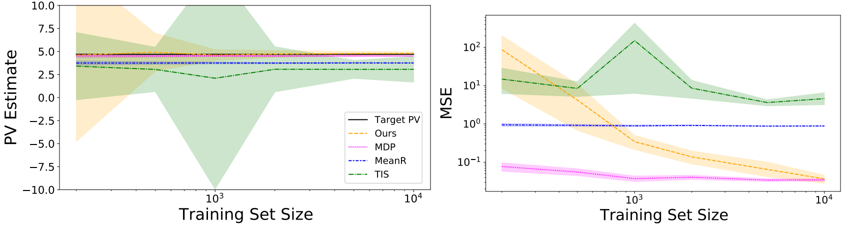


OPE in Confounded POMDP

- Reduces to a sequence of nested proximal causal inference problems
- Subject to certain completeness assumptions analogous to proximal causal inference, can do OPE
- Need to fit value bridge function and action bridge function
  - Analogous to q-function and density ratio
  - Given by conditional moment equations
  - Solve using VMM

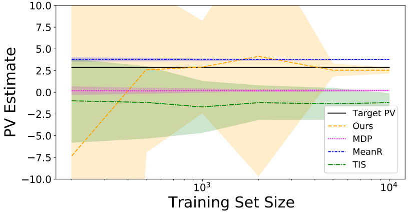
Experiments

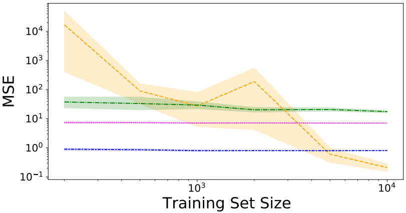
- Easy (high overlap) policy



Experiments

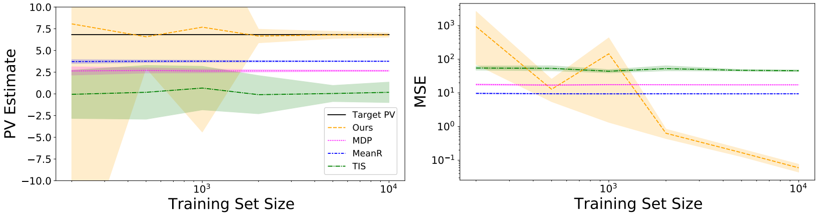
- Hard (low overlap) policy





Experiments

- Optimal policy



Conclusions

- Conditional moment model can be used for many problems
  - Workhorse of economics
  - Important in offline RL
  - Especially confounded settings
  - Can even be used to do cost-sensitive classification
- Sieves are unwieldy $\longrightarrow$ more ML-ish minimax approaches
  - Loses the efficiency and inference of OWGMM 🥇
- Developed VMM by more directly generalizing OWGMM to minimax setting with general function classes
  - Asymptotically normal
  - Semiparametrically efficient
  - Can be applied to itself to estimate standard errors
- Works well in practice
  - ... even beyond finite dim θ 🤔

Thank you!