使用内核方法放宽因果推断中的可观测性假设

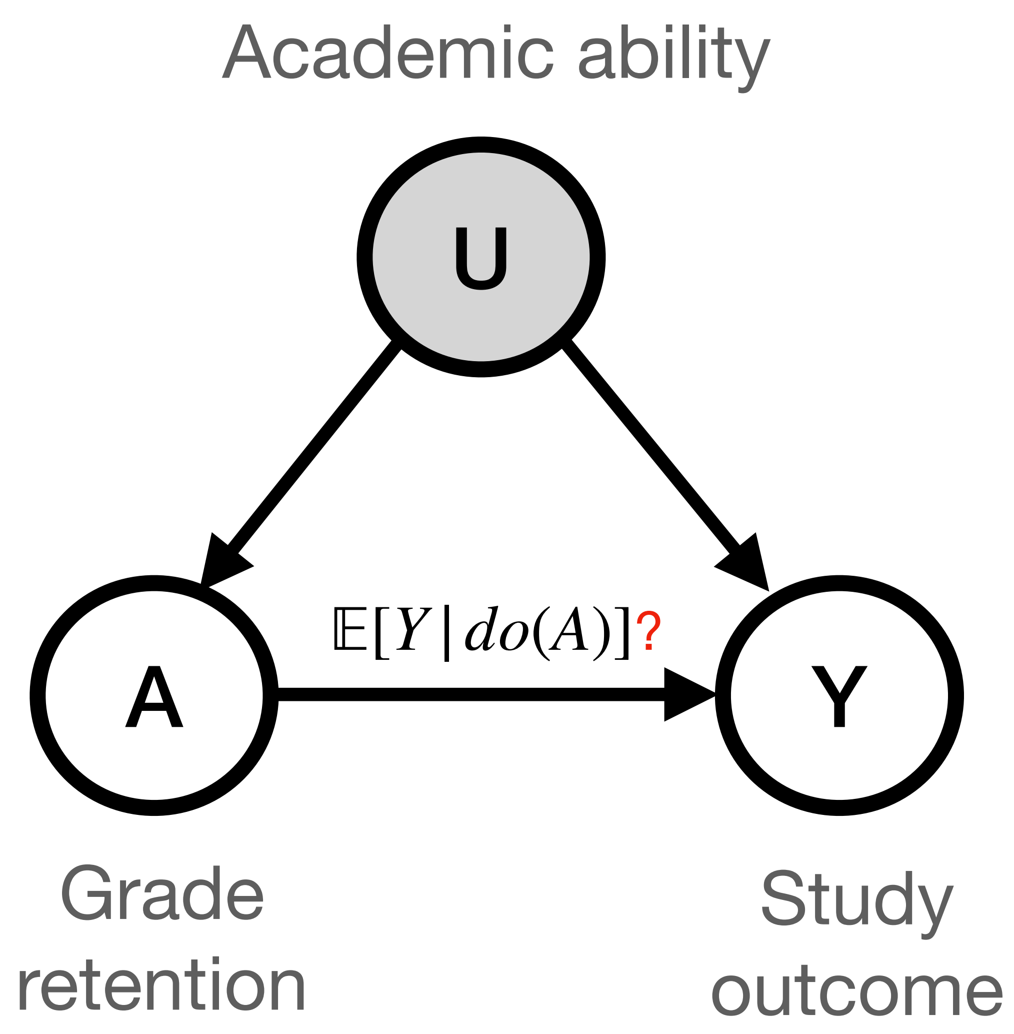
Yuchen Zhu

with Limor Gultchin, Arthur Gretton, Anna Korba, Matt Kusner, Afsaneh Mastouri, Krikamol Muandet, Ricardo Silva

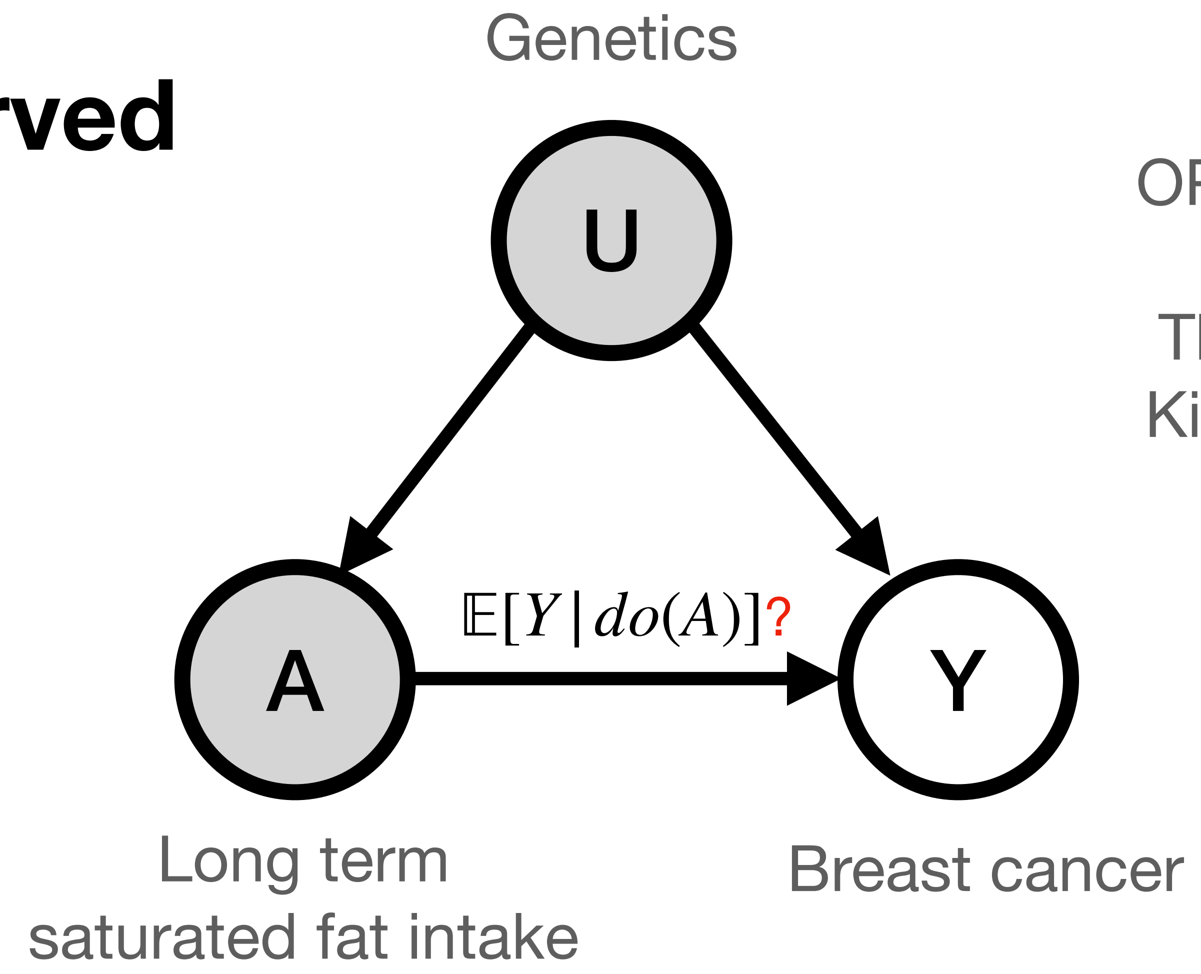
Talk at *Causal Inference Seminars*, 04.2023

Why relax observability assumptions?

Unobserved confounders:



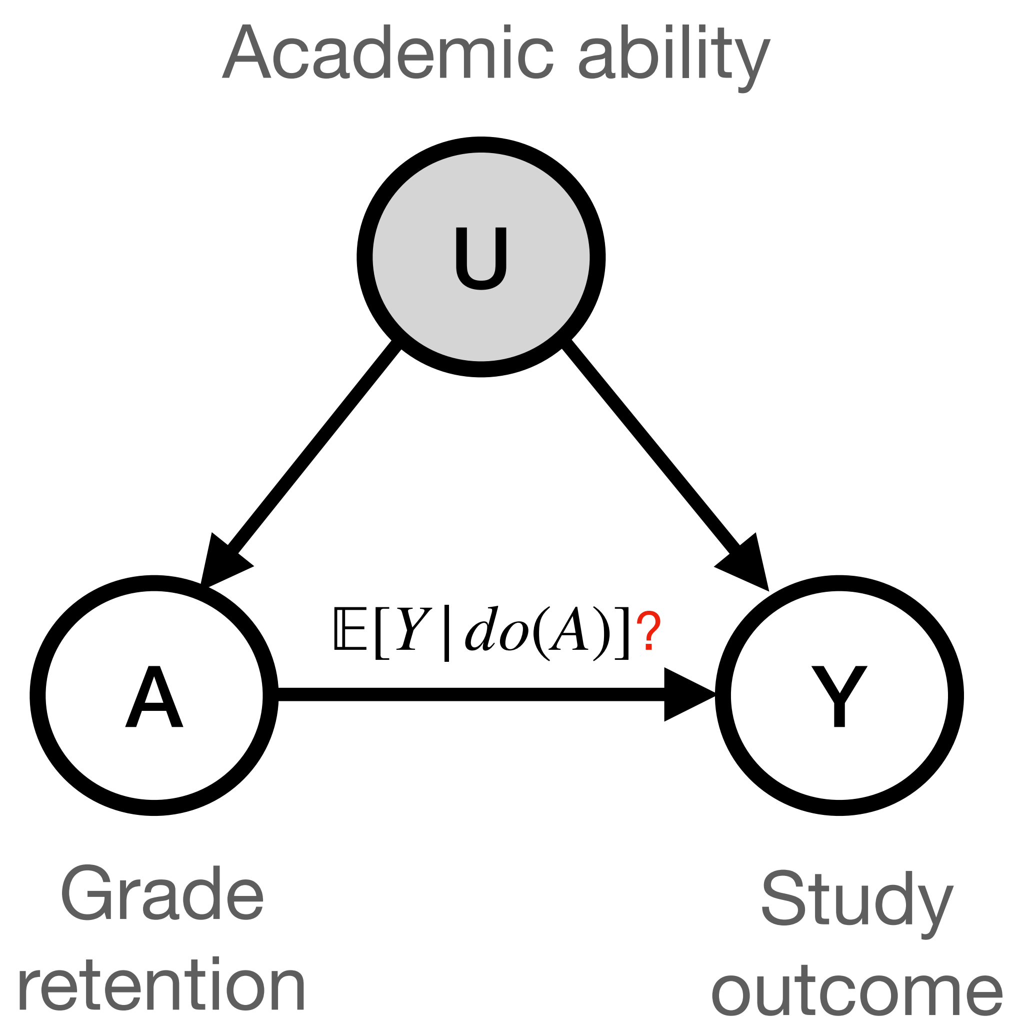
Action observed with error:



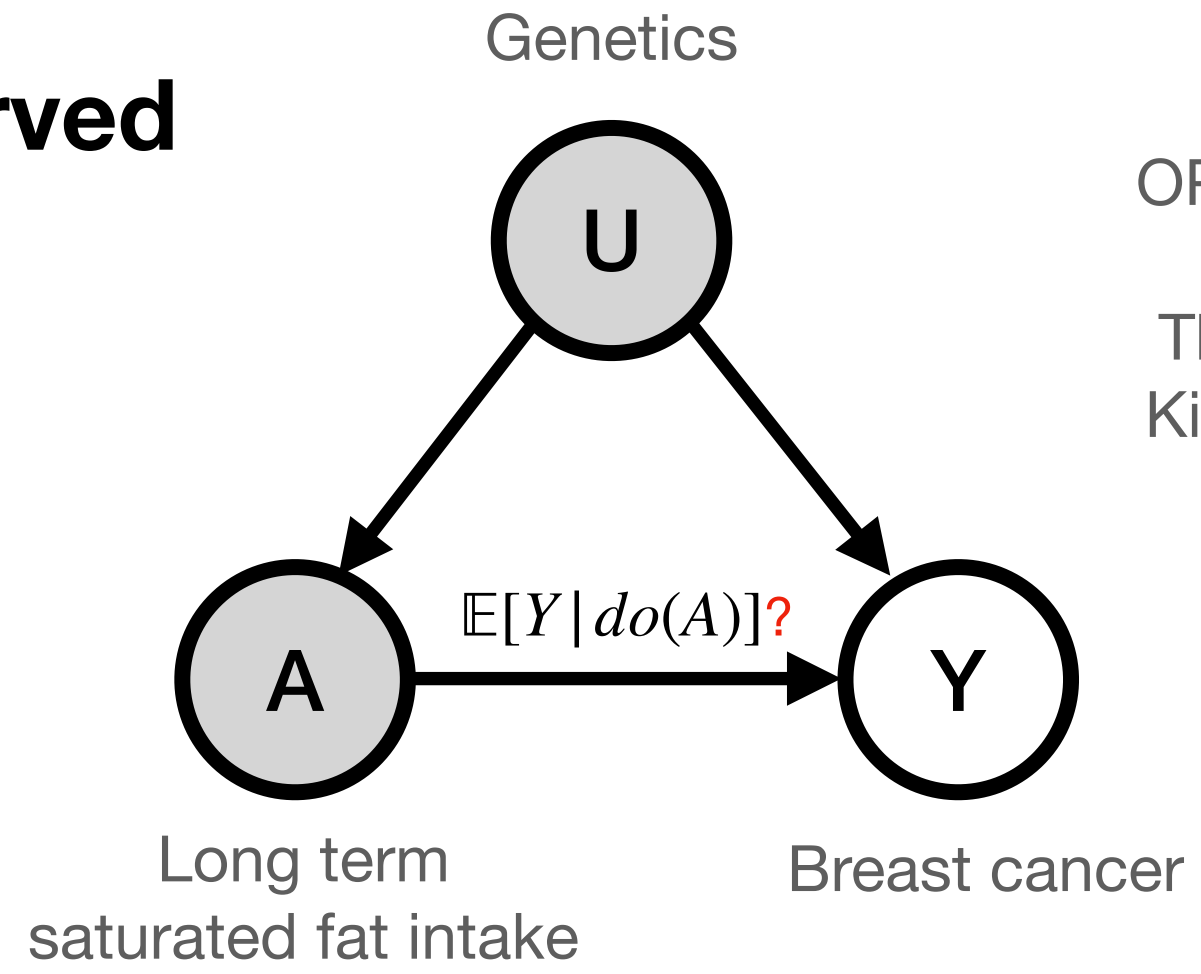
OPEN study: Subar, Thompson, Kipnis, et al. 2001

Why relax observability assumptions?

Unobserved confounders:

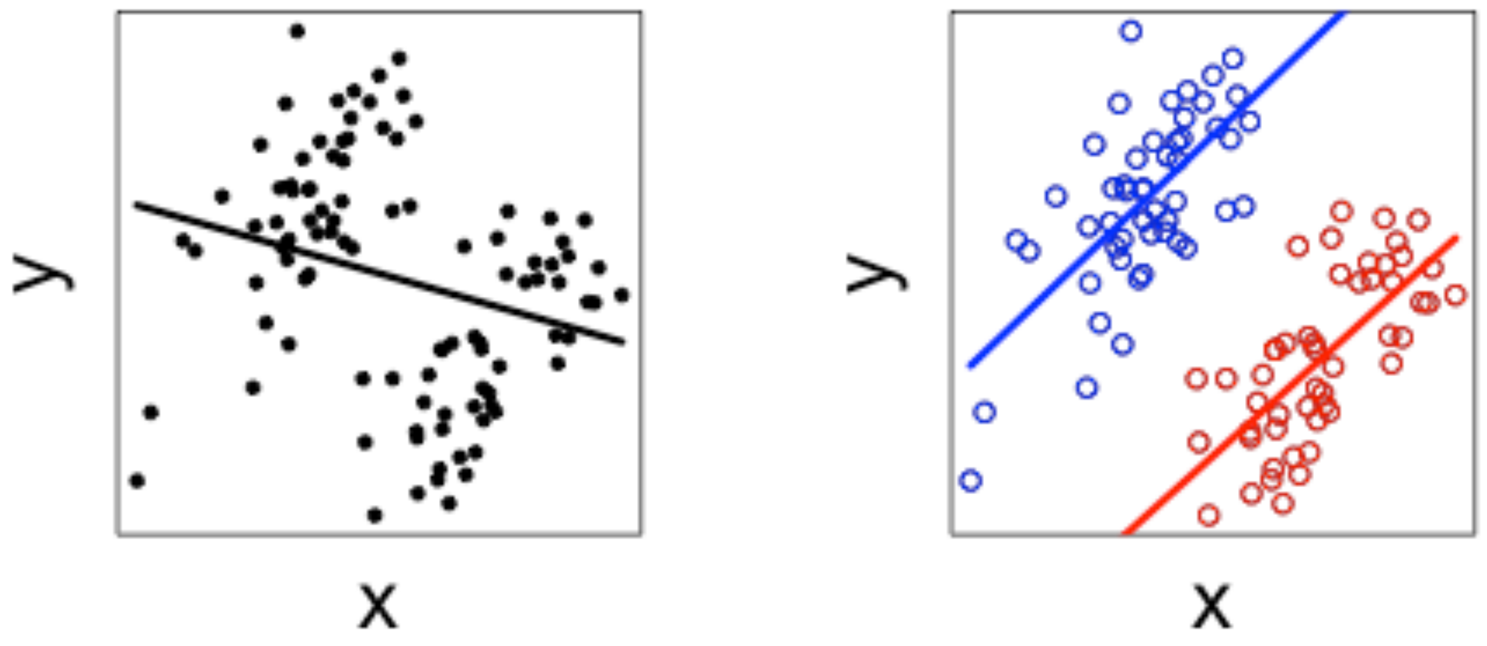


Action observed with error:



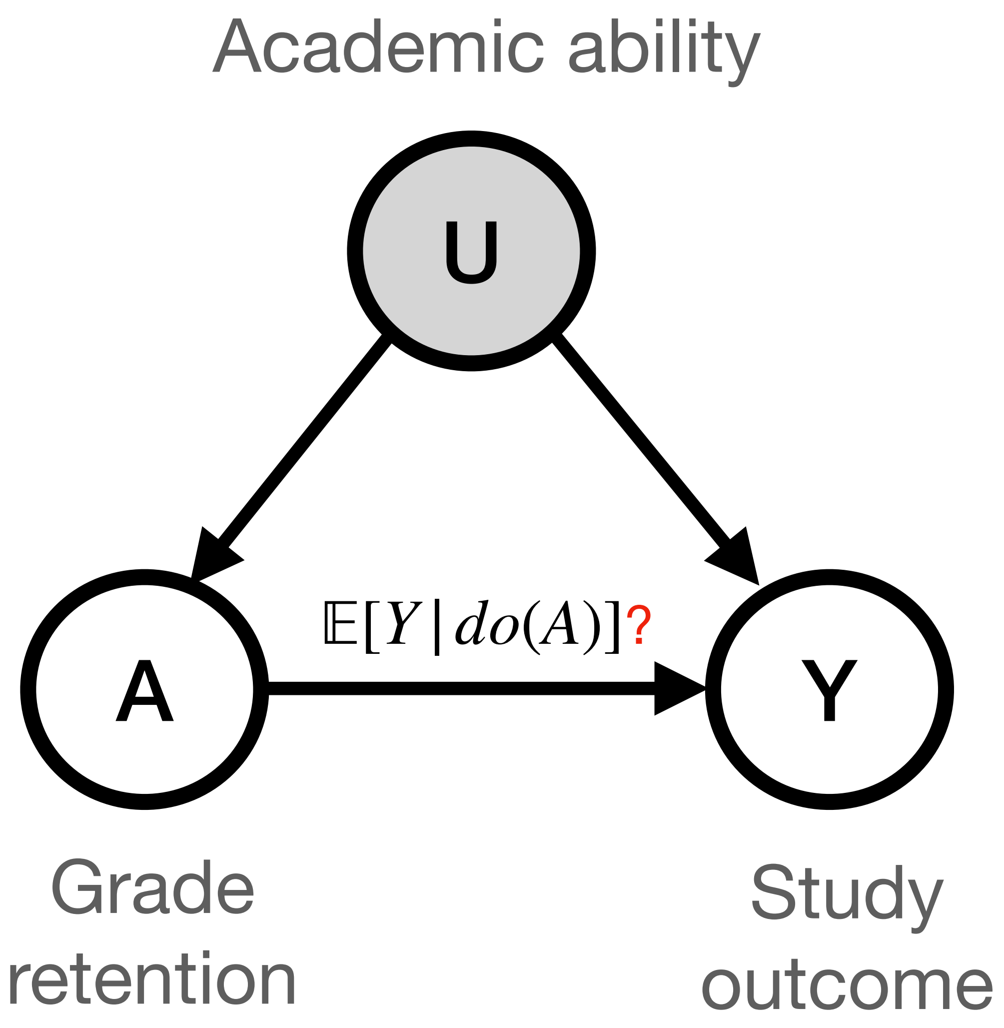
OPEN study: Subar, Thompson, Kipnis, et al. 2001

Simpson's paradox:

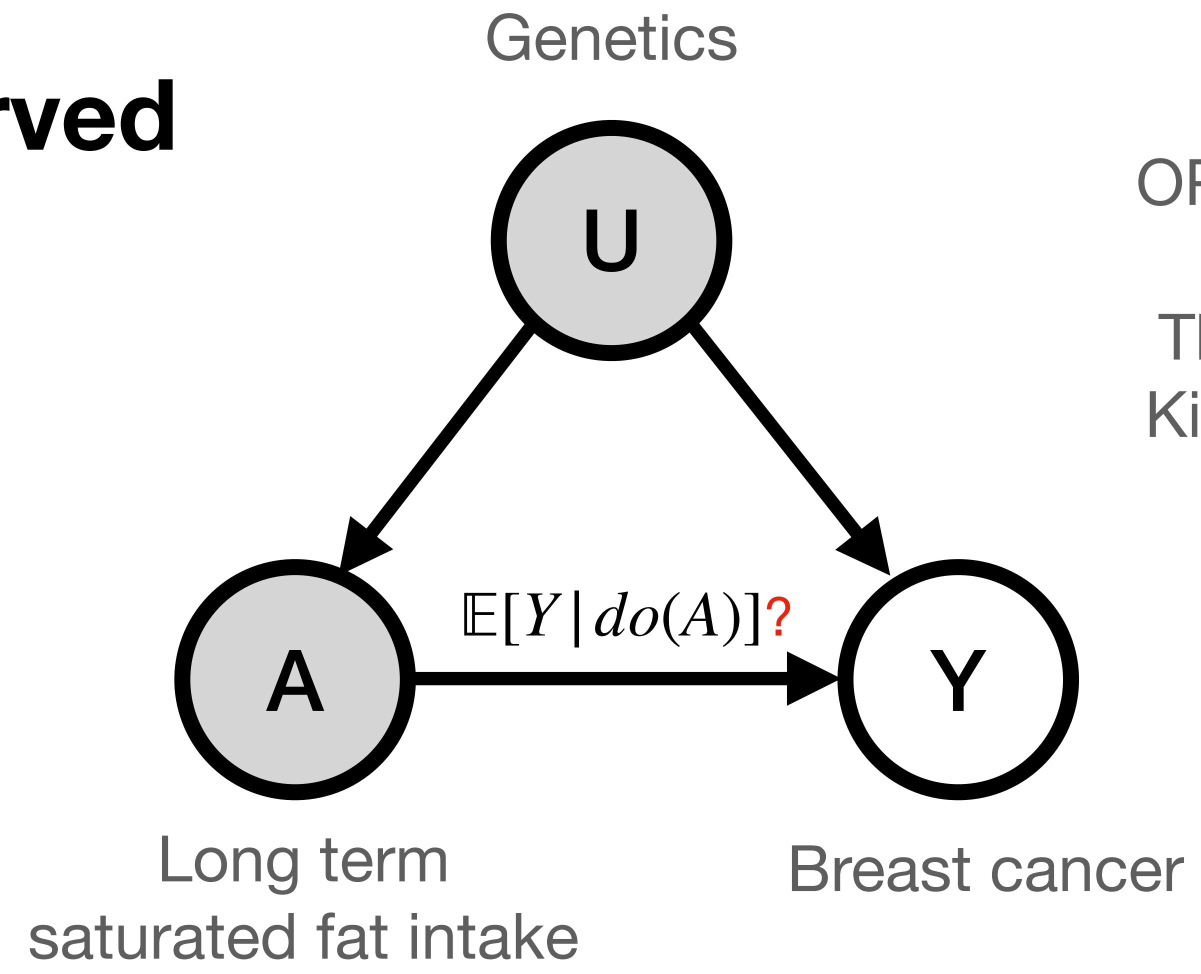


Why relax observability assumptions?

Unobserved confounders:

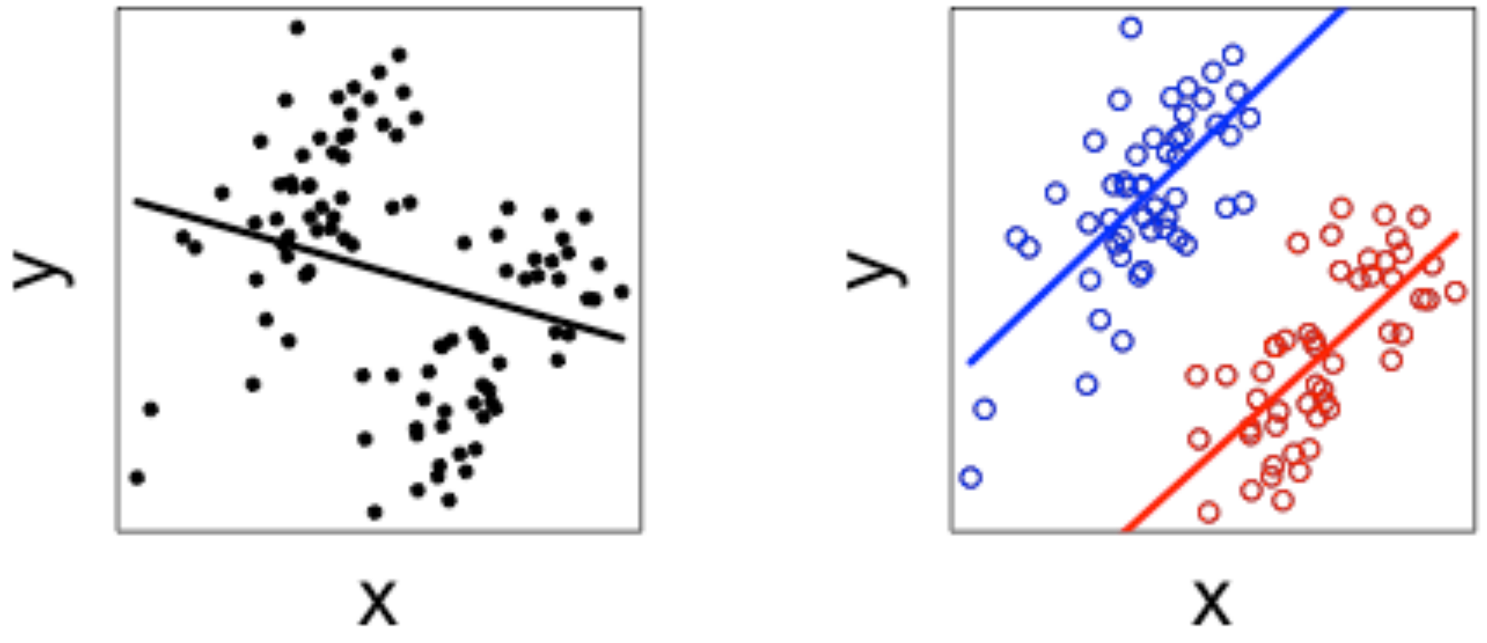


Action observed with error:

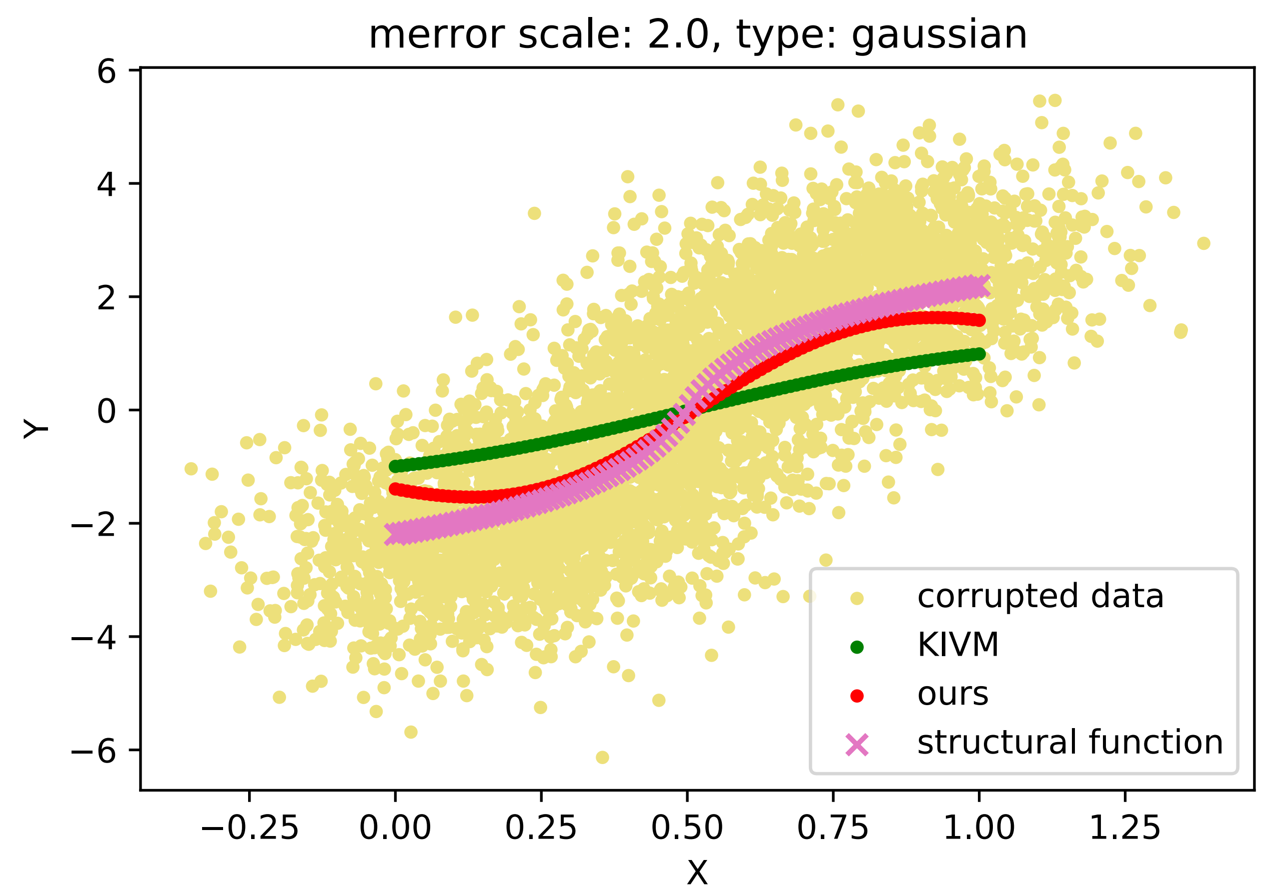


OPEN study: Subar, Thompson, Kipnis, et al. 2001

Simpson's paradox:



Mask interesting relationships:



Kernel Mean Embeddings (KME)

$$\mu_{P_X}(x) = \int k(x, y) P_X(y) dy$$

Kernel Mean Embeddings

$$\mu_{P_X}(x) = \int k(x, y) P_X(y) dy$$

Characteristic kernel: $P_X \xmapsto{\text{Injective}} \mu_{P_X}(y)$

Kernel Mean Embeddings

$$\mu_{P_X}(x) = \int k(x, y) P_X(y) dy$$

Characteristic kernel: $P_X \xmapsto{\text{Injective}} \mu_{P_X}(y)$

$$\langle \mu_{P_X}, f \rangle_{H_X} = \mathbb{E}_{P_X}[f(X)]$$

Conditional Kernel Mean Embeddings (CME)

$$\mu_{W|a,x,z} := C_{W|A,X,Z}\left(\phi(a) \otimes \phi(x) \otimes \phi(z)\right)$$

Conditional Kernel Mean Embeddings (CME)

$$\mu_{W|a,x,z} := C_{W|A,X,Z}\left(\phi(a) \otimes \phi(x) \otimes \phi(z)\right)$$

$$\widehat{C}_{W|A,X,Z} = \underset{C \in \mathcal{H}_\Gamma}{\operatorname{argmin}}\ \widehat{E}(C), \text{ with}$$

$$\widehat{E}(C) = \frac{1}{m}\sum_{i=1}^{m} \|\phi(w_i) - C\phi(a_i, x_i, z_i)\|^2_{\mathcal{H}_\mathcal{W}} + \lambda\ \|C\|^2_{\mathcal{H}_\Gamma}$$

Conditional Kernel Mean Embeddings (CME)

$$\mu_{W|a,x,z} := C_{W|A,X,Z}\left(\phi(a) \otimes \phi(x) \otimes \phi(z)\right)$$

$$\widehat{C}_{W|A,X,Z} = \underset{C \in \mathcal{H}_\Gamma}{\operatorname{argmin}}\ \widehat{E}(C), \text{ with}$$

$$\widehat{E}(C) = \frac{1}{m}\sum_{i=1}^{m} \|\phi(w_i) - C\phi(a_i, x_i, z_i)\|^2_{\mathcal{H}_\mathcal{W}} + \lambda\ \|C\|^2_{\mathcal{H}_\Gamma}$$

$$\widehat{C}_{W|A,X,Z} = \Phi(W)(\mathcal{K}_{AXZ} + m\ \lambda\ )^{-1}\Phi^T(A, X, Z)$$

Conditional Kernel Mean Embeddings (CME)

$$\mu_{W|a,x,z} := C_{W|A,X,Z}\left(\phi(a) \otimes \phi(x) \otimes \phi(z)\right)$$

$$\widehat{C}_{W|A,X,Z} = \underset{C \in \mathcal{H}_\Gamma}{\operatorname{argmin}}\ \widehat{E}(C), \text{ with}$$

$$\widehat{E}(C) = \frac{1}{m}\sum_{i=1}^{m} \|\phi(w_i) - C\phi(a_i, x_i, z_i)\|^2_{\mathcal{H}_\mathcal{W}} + \lambda\ \|C\|^2_{\mathcal{H}_\Gamma}$$

$$\widehat{C}_{W|A,X,Z} = \Phi(W)(\mathcal{K}_{AXZ} + m\ \lambda\ )^{-1}\Phi^T(A, X, Z)$$

Convergence rates are well understood (Singh et al 2019, Mastouri, Zhu, et al 2021)

Connection with Characteristic Functions

Translation invariant: $k(x, y) = k(x - y)$

$$\mu(x) = \int k(x - y)p(y)dy$$

Connection with Characteristic Functions

Translation invariant: $k(x, y) = k(x - y)$

$$\mu(x) = \int k(x - y)p(y)dy$$

$$\hat{\mu}[\alpha] = \hat{k}[\alpha]\psi[\alpha]$$

Connection with Characteristic Functions

Translation invariant: $k(x, y) = k(x - y)$

$$\mu(x) = \int k(x - y)p(y)dy$$

$$\hat{\mu}[\alpha] = \hat{k}[\alpha]\psi[\alpha]$$

Bochner's theorem: $\hat{k}$ is a probability measure.

Connection with Characteristic Functions

KRR estimate of CME:

$$\hat{\mu}^{(s)}_{X|z}(x) = \sum_{j=1}^{s} \hat{\gamma}^{(s)}_j(z)k(x_j, x)$$

$$\hat{\gamma}^{(s)}_j(z) = (K_Z + s\lambda I)^{-1}K_{Zz}$$

Connection with Characteristic Functions

KRR estimate of CME:

$$\hat{\mu}^{(s)}_{X|z}(x) = \sum_{j=1}^{s} \hat{\gamma}^{(s)}_j(z) k(x_j, x)$$

$$\hat{\gamma}^{(s)}_j(z) = (K_Z + s\lambda I)^{-1} K_{Zz}$$

Fourier transform:

$$\tilde{\hat{\mu}}^{(s)}_{X|z}(\alpha) = \sum_{j=1}^{s} \hat{\gamma}^{(s)}_j(z) e^{-i\alpha x_j} \tilde{k}(\alpha)$$

Connection with Characteristic Functions

KRR estimate of CME:

$$\hat{\mu}^{(s)}_{X|z}(x) = \sum_{j=1}^{s} \hat{\gamma}^{(s)}_j(z) k(x_j, x)$$

$$\hat{\gamma}^{(s)}_j(z) = (K_Z + s\lambda I)^{-1} K_{Zz}$$

Fourier transform:

$$\tilde{\hat{\mu}}^{(s)}_{X|z}(\alpha) = \sum_{j=1}^{s} \hat{\gamma}^{(s)}_j(z) e^{-i\alpha x_j}\, \tilde{k}(\alpha)$$

$$= \tilde{k}(\alpha) \underbrace{\sum_{j=1}^{s} \hat{\gamma}^{(s)}_j(z) e^{-j\alpha x_j}}_{=:\hat{\psi}^{(s)}_{\mathcal{P}_{X|z}}(-\alpha)}$$

Connection with Characteristic Functions

$(x_j, z_j)_{j=1}^{s}$

Connection with Characteristic Functions

$(x_j, z_j)_{j=1}^{s}$ $\longrightarrow$ Have $\hat{\mu}^n_{X|z}(y) = \sum_{j=1}^{n} \hat{\gamma}_j^n(z) k(x_j, y)$.

Where $\hat{\gamma}_j^n(z) = (K_{ZZ} + n\hat{\lambda}^n I)^{-1} K_{Zz}$.

Connection with Characteristic Functions

$(x_j, z_j)_{j=1}^{s}$ $\longrightarrow$ Have $\hat{\mu}^n_{X|z}(y) = \sum_{j=1}^{n} \hat{\gamma}^n_j(z) k(x_j, y)$.

Let $\hat{\psi}^n_{X|z}(\alpha) := \sum_{j=1}^{n} \hat{\gamma}^n_j(z) e^{i\alpha x_j}$.

Where $\hat{\gamma}^n_j(z) = (K_{ZZ} + n\hat{\lambda}^n I)^{-1} K_{Zz}$.

Connection with Characteristic Functions

$(x_j, z_j)_{j=1}^{s}$ $\longrightarrow$ Have $\hat{\mu}^n_{X|z}(y) = \sum_{j=1}^{n} \hat{\gamma}^n_j(z) k(x_j, y).$

Let $\hat{\psi}^n_{X|z}(\alpha) := \sum_{j=1}^{n} \hat{\gamma}^n_j(z) e^{i\alpha x_j}.$

Where $\hat{\gamma}^n_j(z) = (K_{ZZ} + n\hat{\lambda}^n I)^{-1} K_{Zz}.$

Theorem 1. With real, translation-invariant kernel:
$\hat{\mu}^n_{X|Z} \to^n \mu_{X|Z}$ iff $\hat{\psi}^n_{X|Z} \to^n \psi_{X|Z}$ in IFT of kernel.

Zhu et al 2022, UAI.

Kotlarski's Lemma

LEMMA 1. *Let* X_1, X_2, X_3 *be three independent real random variables*

Kotlarski's Lemma

LEMMA 1. *Let X_1, X_2, X_3 be three independent real random variables, and let*

$$Z_1 = X_1 - X_3, Z_2 = X_2 - X_3 .$$

Kotlarski's Lemma

LEMMA 1. *Let* X_1, X_2, X_3 *be three independent real random variables, and let*

$$Z_1 = X_1 - X_3, Z_2 = X_2 - X_3 .$$

If the characteristic function of the pair (Z_1, Z_2) *does not vanish,*

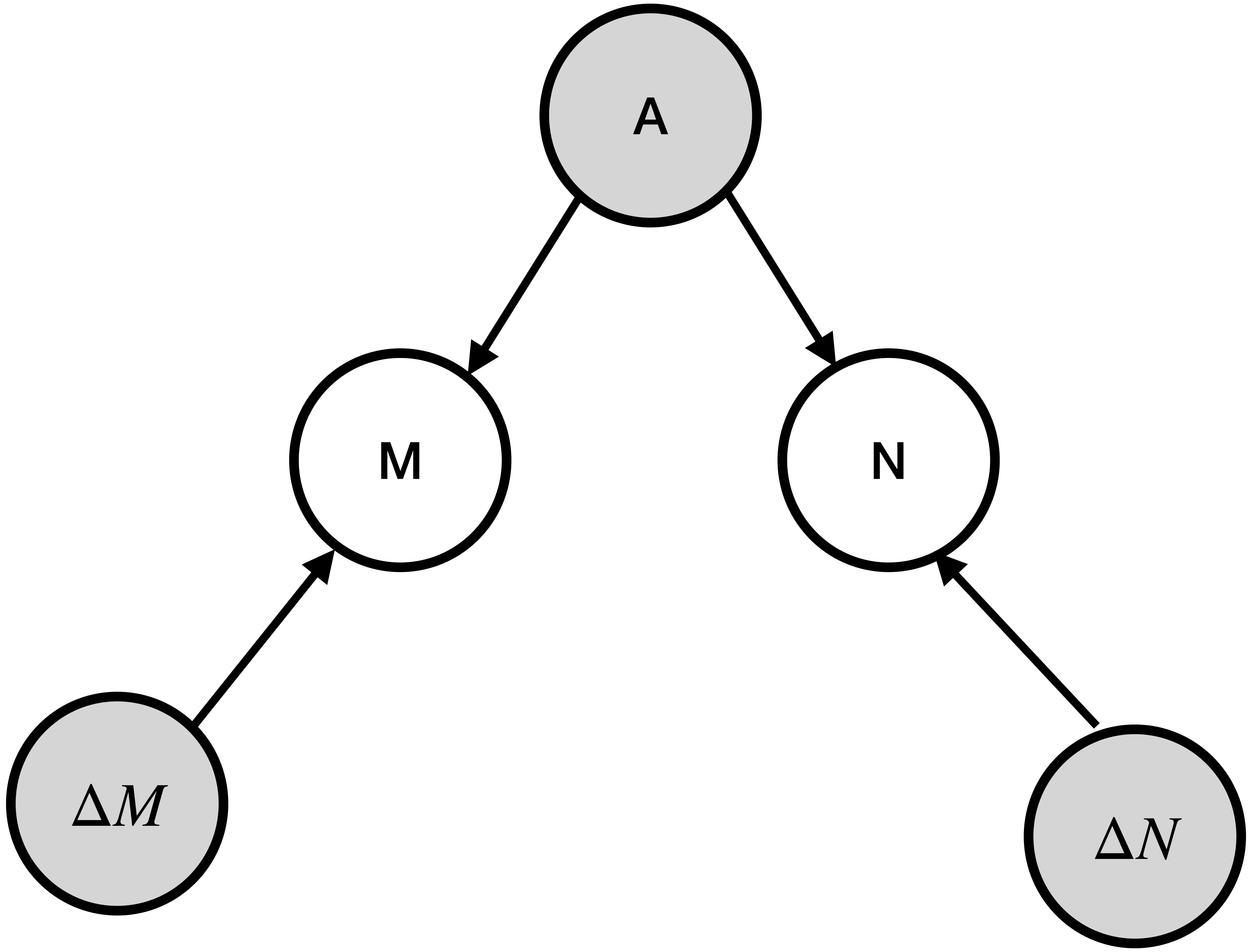
Kotlarski's Lemma

LEMMA 1. *Let X_1, X_2, X_3 be three independent real random variables, and let*

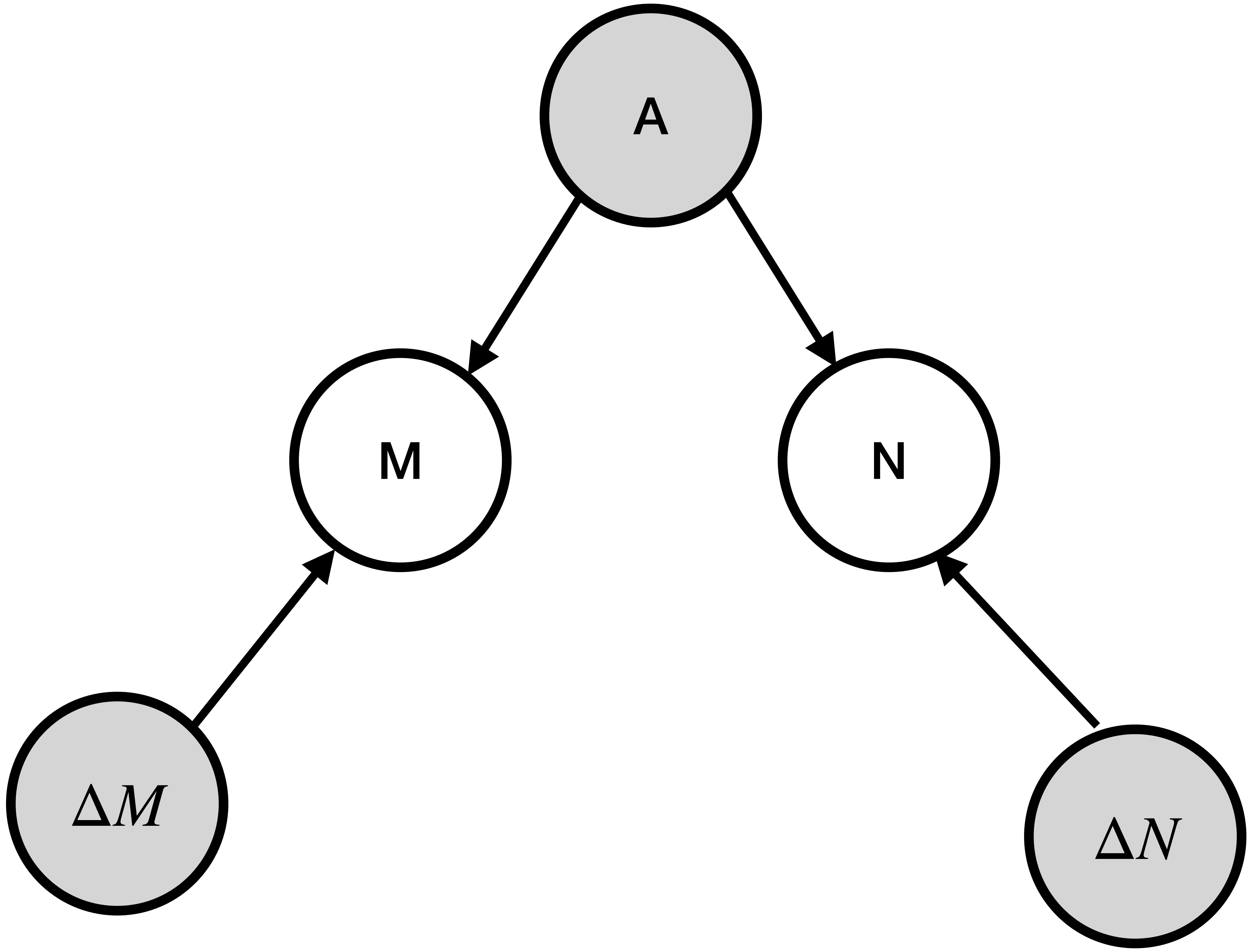
$$Z_1 = X_1 - X_3, Z_2 = X_2 - X_3 .$$

If the characteristic function of the pair (Z_1, Z_2) does not vanish, then the distribution of (Z_1, Z_2) determines the distributions of X_1, X_2, X_3 up to a change of the location.

Kotlarski's Lemma



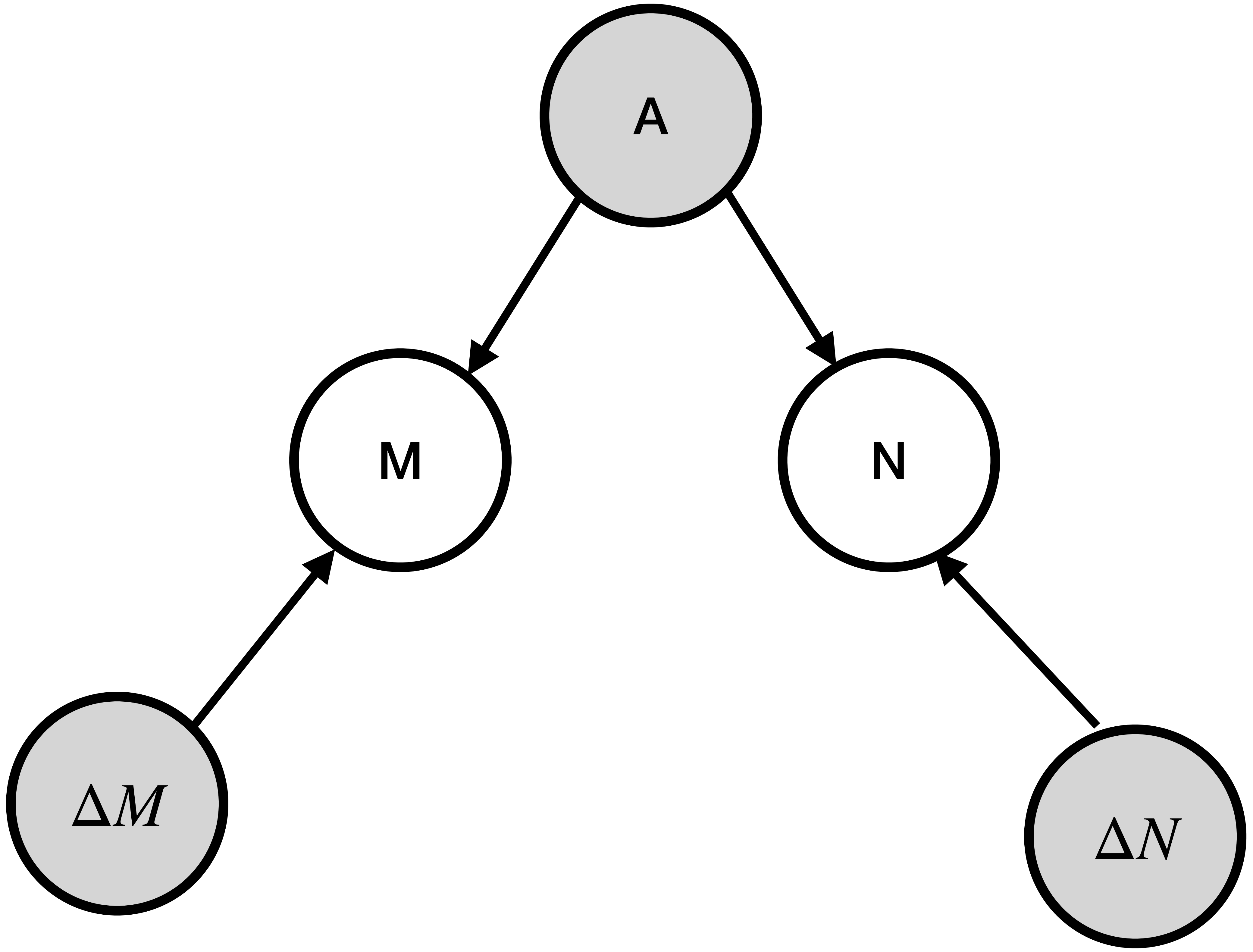
Kotlarski's Lemma



$$M = A + \Delta M$$

$$N = A + \Delta N$$

Kotlarski's Lemma

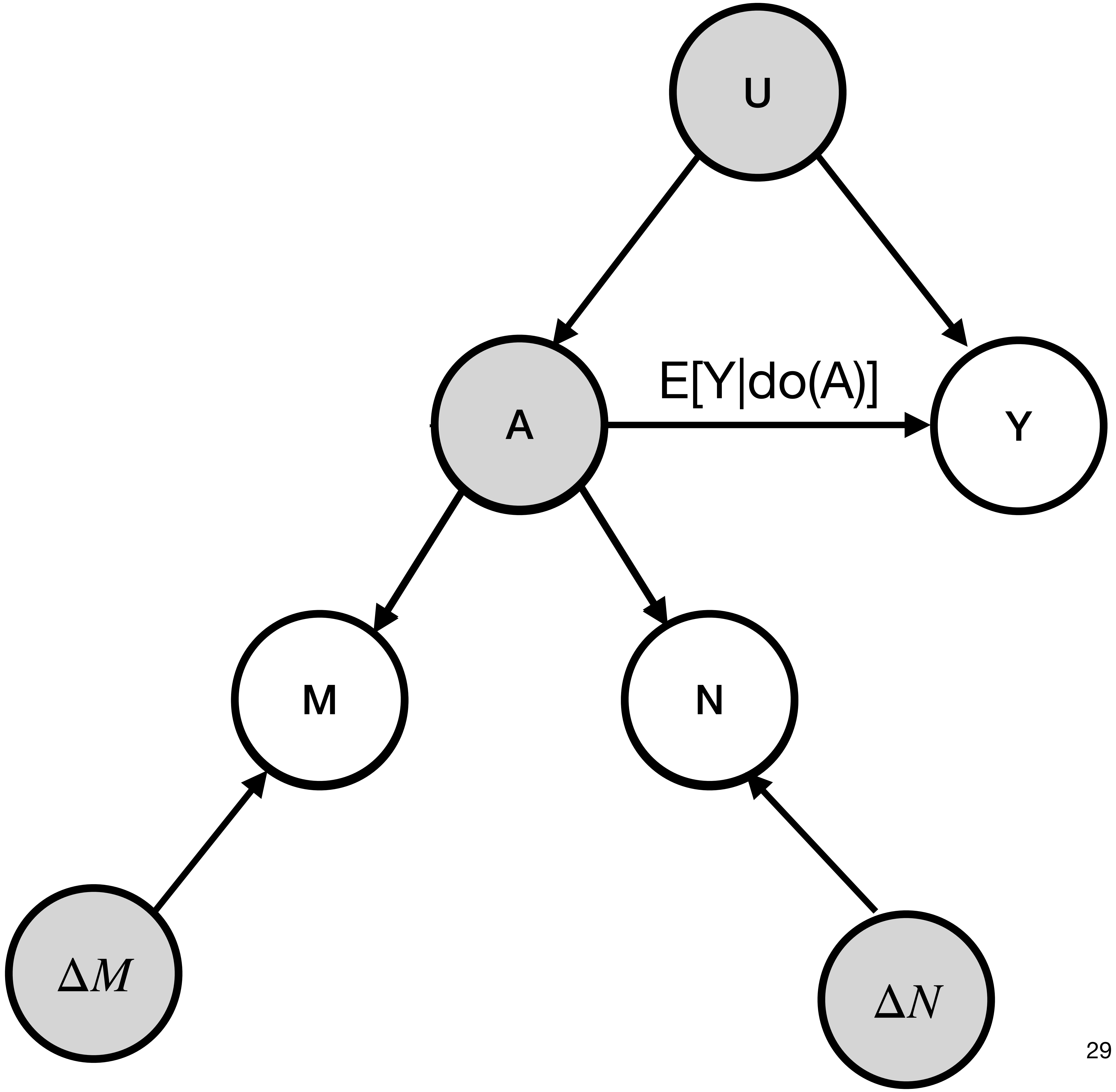


$$M = A + \Delta M$$

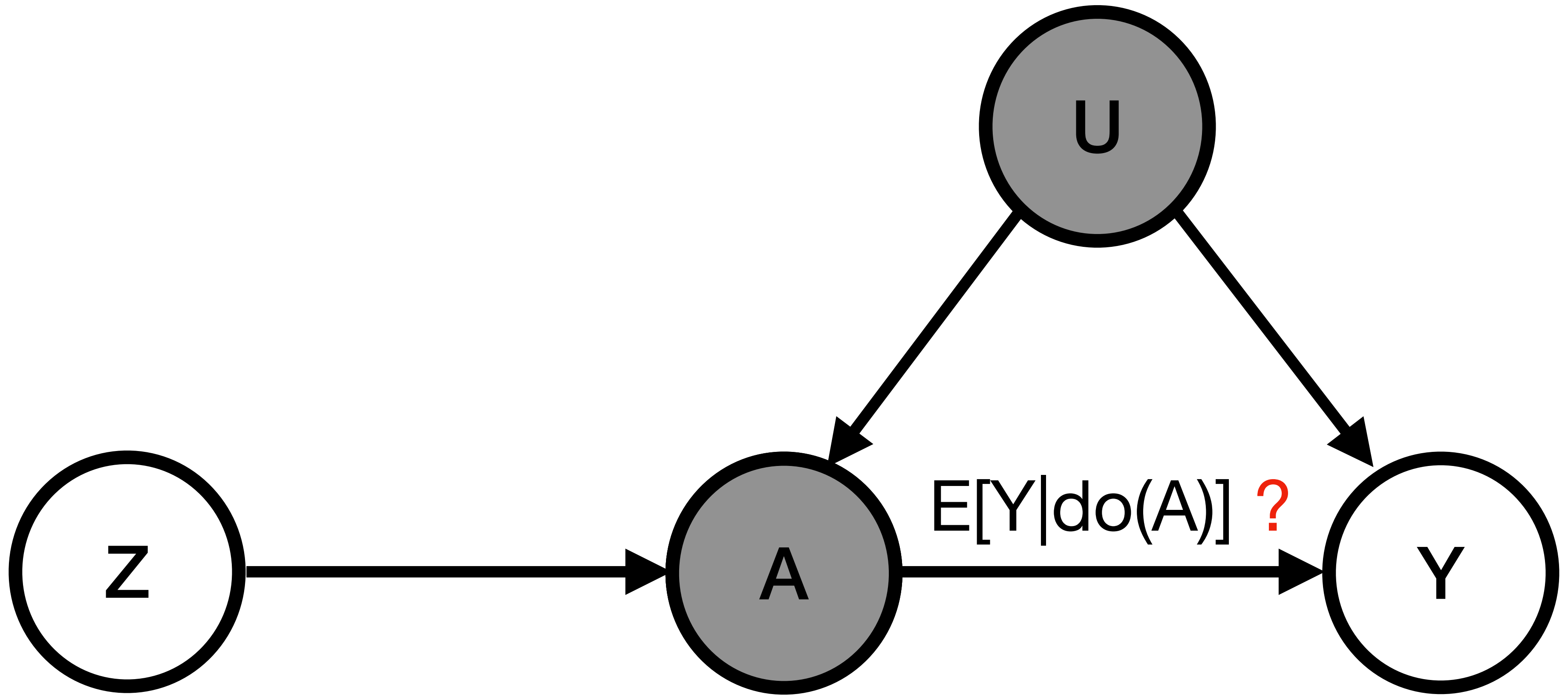
$$N = A + \Delta N$$

$$\overbrace{\mathbb{E}_{\mathscr{P}_A}\left[e^{i\alpha A}\right]}^{\psi_{\mathscr{P}_A}(\alpha):=} = \exp\left(\int_0^\alpha i\frac{\mathbb{E}\left[Me^{i\nu N}\right]}{\mathbb{E}\left[e^{i\nu N}\right]}d\nu\right)$$

Application in causal inference with corrupted treatments



Recap: Identification with instrumental variables



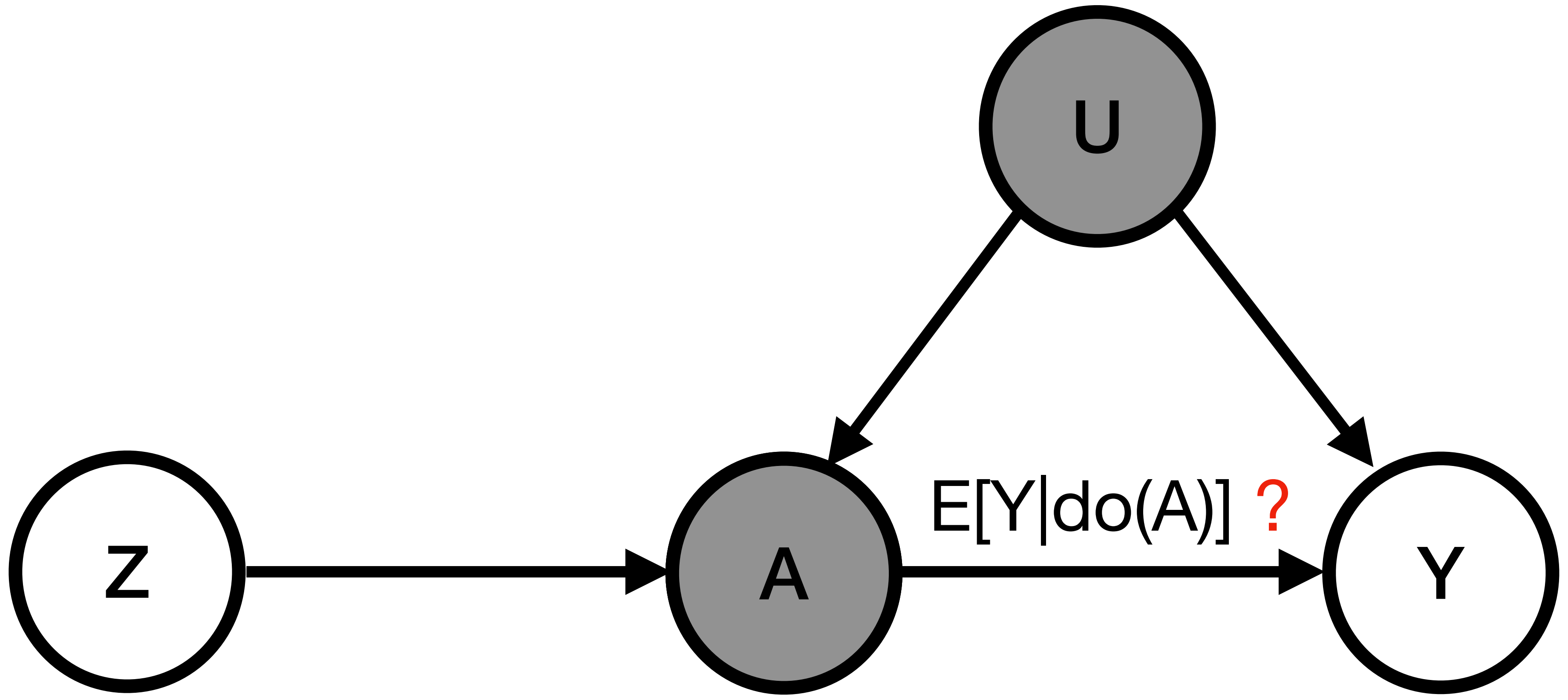
Identification:

$$Y = f(A) + \epsilon \quad \mathbb{E}[\epsilon \,|\, Z] = 0$$

$$f(A) = \mathbb{E}[Y \,|\, do(A)]$$

$$\mathbb{E}[Y \,|\, Z] = \int_{\mathscr{A}} f(a) p(a \,|\, Z) da$$

Recap: Identification with instrumental variables



Identification:

$$Y = f(A) + \epsilon \quad \mathbb{E}[\epsilon \,|\, Z] = 0$$

$$f(A) = \mathbb{E}[Y \,|\, do(A)]$$

$$\mathbb{E}[Y \,|\, Z] = \int_{\mathscr{A}} f(a)p(a \,|\, Z)da$$

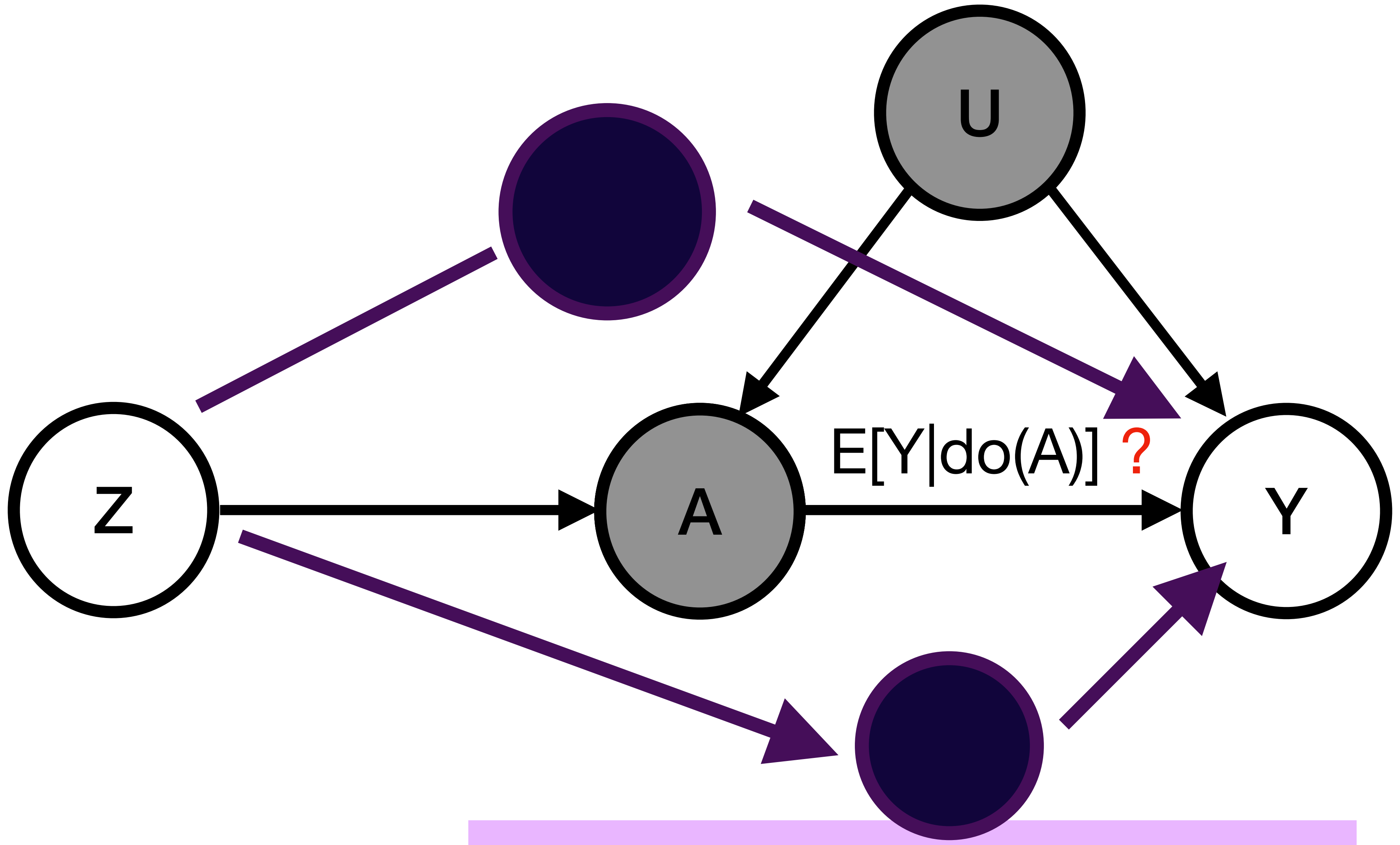
Linear case:

$$Y = \beta A + \epsilon_Y \quad \epsilon_Y \perp Z$$

$$A = \gamma Z + \epsilon_A \quad \epsilon_A \perp Z$$

$$\implies Y = \beta\gamma Z + \beta\epsilon_A + \epsilon_Y$$

Recap: Identification with instrumental variables



(Strong) Assumptions:
- **Additive error model**
- $(Z! \perp A)_G$
- $(Z \perp Y)_{G_{\bar{A}}}$

Identification:

$$Y = f(A) + \epsilon \quad \mathbb{E}[\epsilon \,|\, Z] = 0$$

$$f(A) = \mathbb{E}[Y \,|\, do(A)]$$

$$\mathbb{E}[Y \,|\, Z] = \int_{\mathscr{A}} f(a)p(a \,|\, Z)da$$

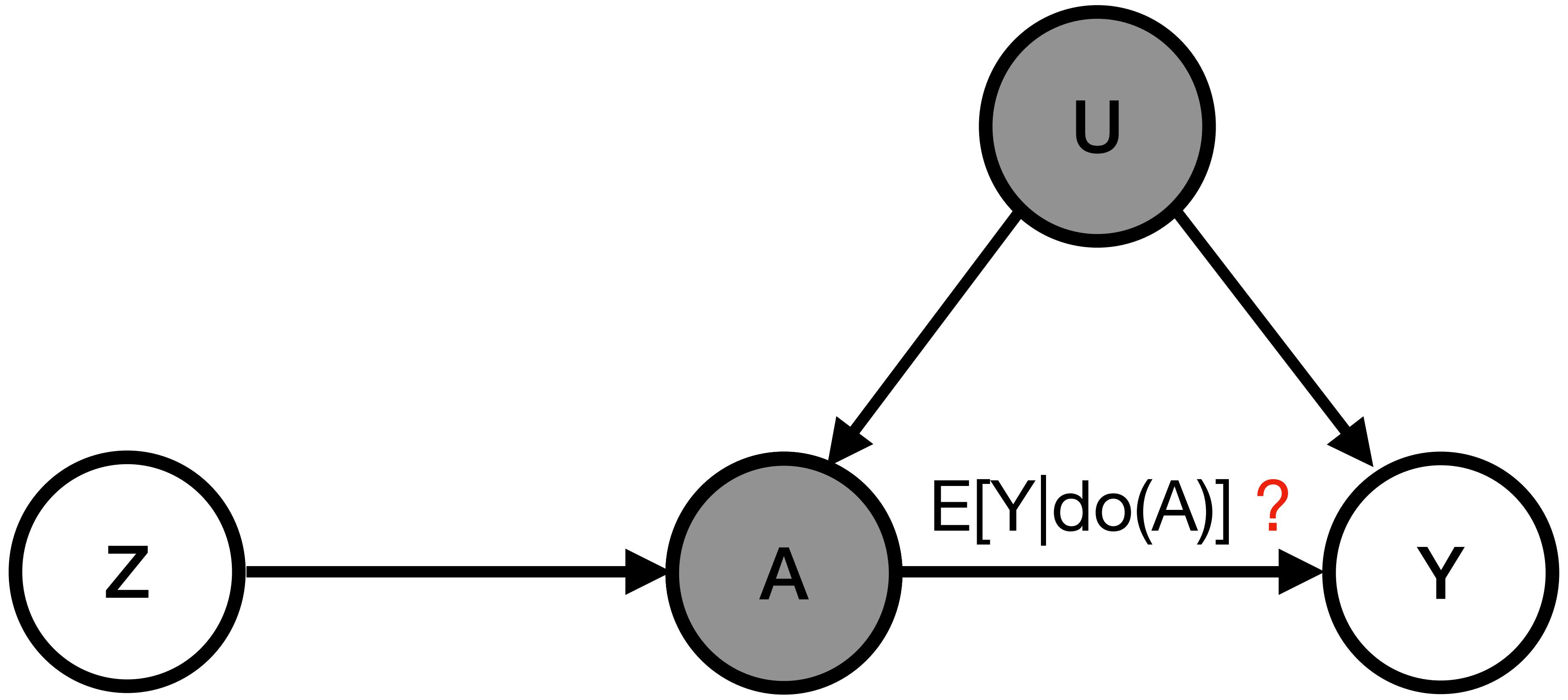
Linear case:

$$Y = \beta A + \epsilon_Y \quad \epsilon_Y \perp Z$$

$$A = \gamma Z + \epsilon_A \quad \epsilon_A \perp Z$$

$$\implies Y = \beta\gamma Z + \beta\epsilon_A + \epsilon_Y$$

Recap: Identification with instrumental variables



Identification:

$$Y = f(A) + \epsilon \quad \mathbb{E}[\epsilon \,|\, Z] = 0$$

$$f(A) = \mathbb{E}[Y \,|\, do(A)]$$

$$\mathbb{E}[Y \,|\, Z] = \int_{\mathscr{A}} f(a)p(a \,|\, Z)da$$

Linear case:

$$Y = \beta A + \epsilon_Y \quad \epsilon_Y \perp Z$$

$$A = \gamma Z + \epsilon_A \quad \epsilon_A \perp Z$$

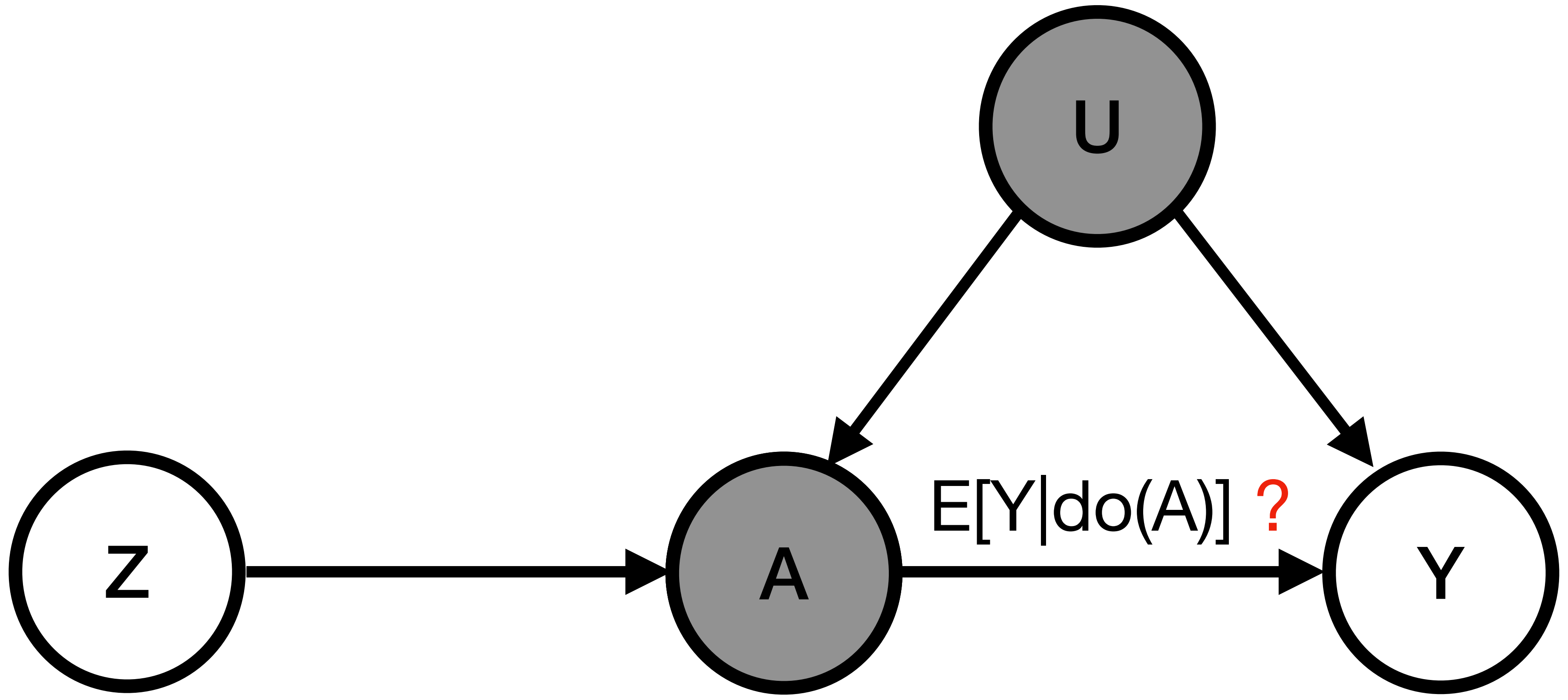
$$\Longrightarrow \; Y = \beta\gamma Z + \beta\epsilon_A + \epsilon_Y$$

(Strong) Assumptions:

- **Additive error model**
- $(Z ! \perp A)_G$
- $(Z \perp Y)_{G_{\bar{A}}}$

False IV: using same 'IV' for several different actions.

Recap: Identification with instrumental variables



Identification:

$$Y = f(A) + \epsilon \quad \mathbb{E}[\epsilon | Z] = 0$$

$$f(A) = \mathbb{E}[Y | do(A)]$$

$$\mathbb{E}[Y|Z] = \int_{\mathscr{A}} f(a)p(a|Z)da$$

???

But if $f(a) = \theta^T \phi(a)$, then simplies to

$$\mathbb{E}[Y|Z] = \theta^T \mathbb{E}[\phi(A)|Z]$$

Linear case:

$$Y = \beta A + \epsilon_Y \quad \epsilon_Y \perp Z$$

$$A = \gamma Z + \epsilon_A \quad \epsilon_A \perp Z$$

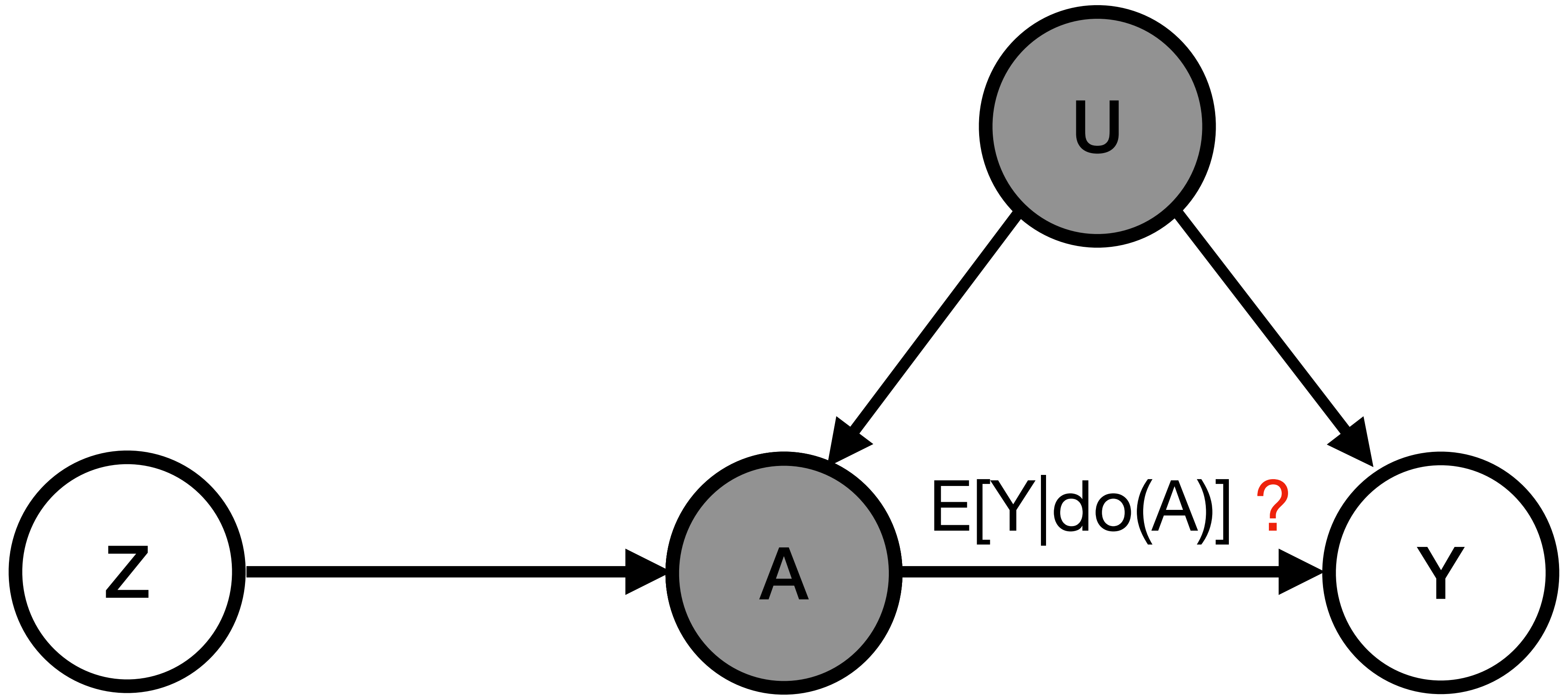
$$\implies Y = \beta\gamma Z + \beta\epsilon_A + \epsilon_Y$$

(Strong) Assumptions:

- **Additive error model**
- $(Z ! \perp A)_G$
- $(Z \perp Y)_{G_{\bar{A}}}$

False IV: using same 'IV' for several different actions.

Recap: Identification with instrumental variables



Identification:

$$Y = f(A) + \epsilon \quad \mathbb{E}[\epsilon \mid Z] = 0$$

$$f(A) = \mathbb{E}[Y \mid do(A)]$$

$$\mathbb{E}[Y \mid Z] = \int_{\mathscr{A}} f(a) p(a \mid Z) da$$

???

But if $f(a) = \theta^T \phi(a)$, then simplies to

$$\mathbb{E}[Y \mid Z] = \theta^T \mathbb{E}[\phi(A) \mid Z]$$

Linear case:

$$Y = \beta A + \epsilon_Y \quad \epsilon_Y \perp Z$$

$$A = \gamma Z + \epsilon_A \quad \epsilon_A \perp Z$$

$$\implies Y = \beta \gamma Z + \beta \epsilon_A + \epsilon_Y$$

(Strong) Assumptions:

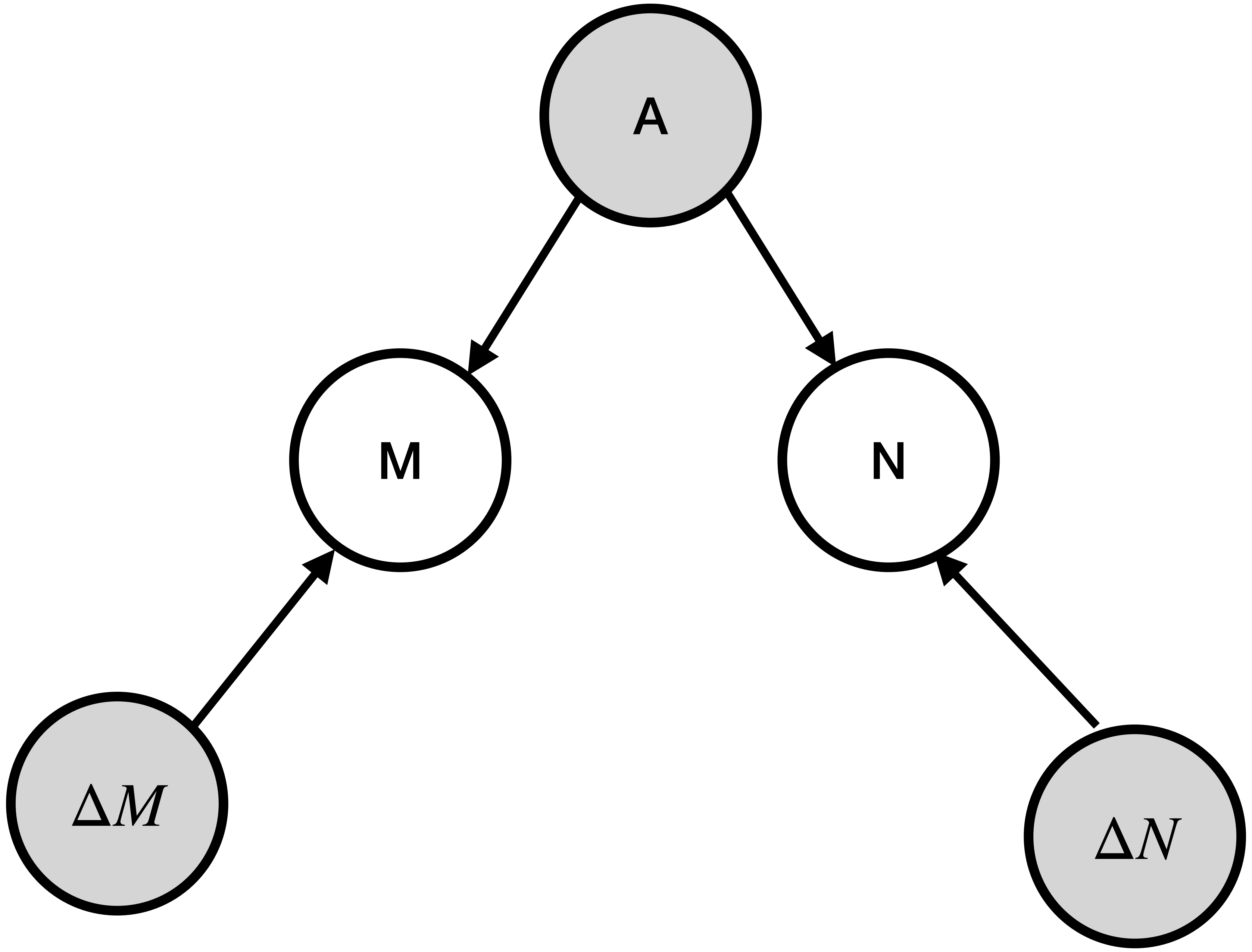
- **Additive error model**
- $(Z \not\perp A)_G$
- $(Z \perp Y)_{G_{\bar{A}}}$

To induce well-posedness:

- Assume f in RKHS.
- Tikhonov regularisation.

False IV: using same 'IV' for several different actions.

Kotlarski's Lemma

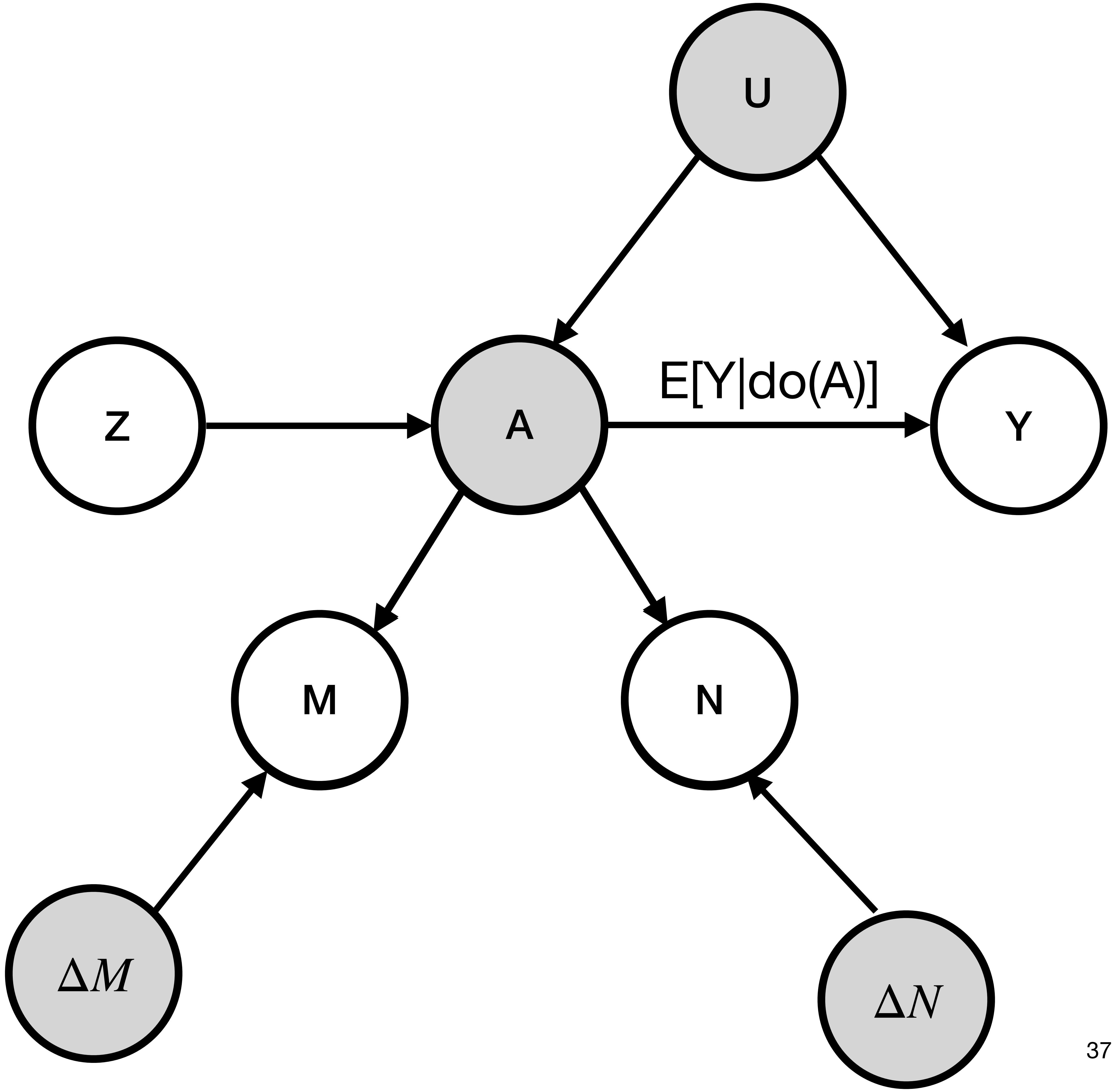


$$M = A + \Delta M$$

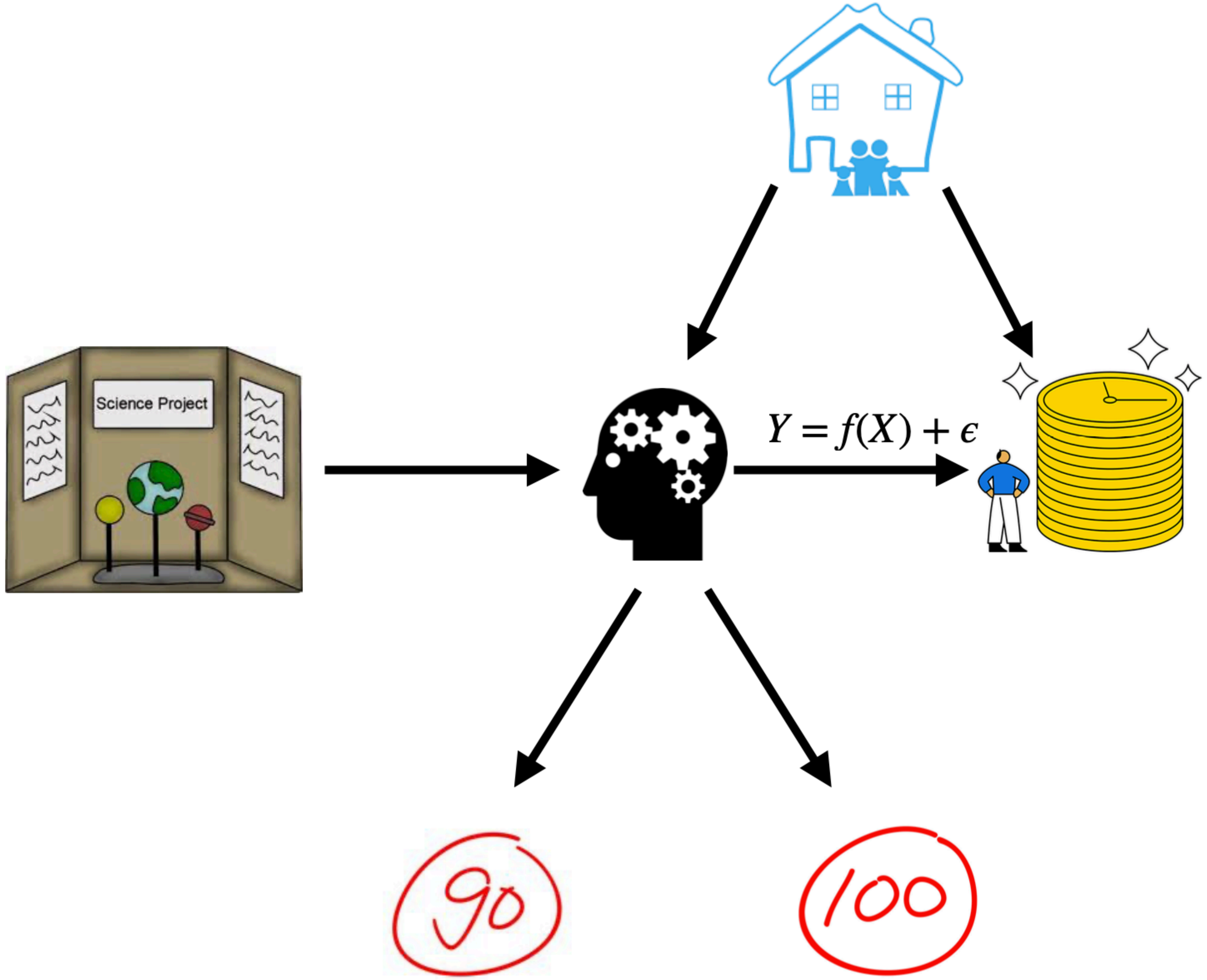
$$N = A + \Delta N$$

$$\overbrace{\mathbb{E}_{\mathscr{P}_A}\left[e^{i\alpha A}\right]}^{\psi_{\mathscr{P}_A}(\alpha):=} = \exp\left(\int_0^\alpha i\frac{\mathbb{E}\left[Me^{i\nu N}\right]}{\mathbb{E}\left[e^{i\nu N}\right]}d\nu\right)$$

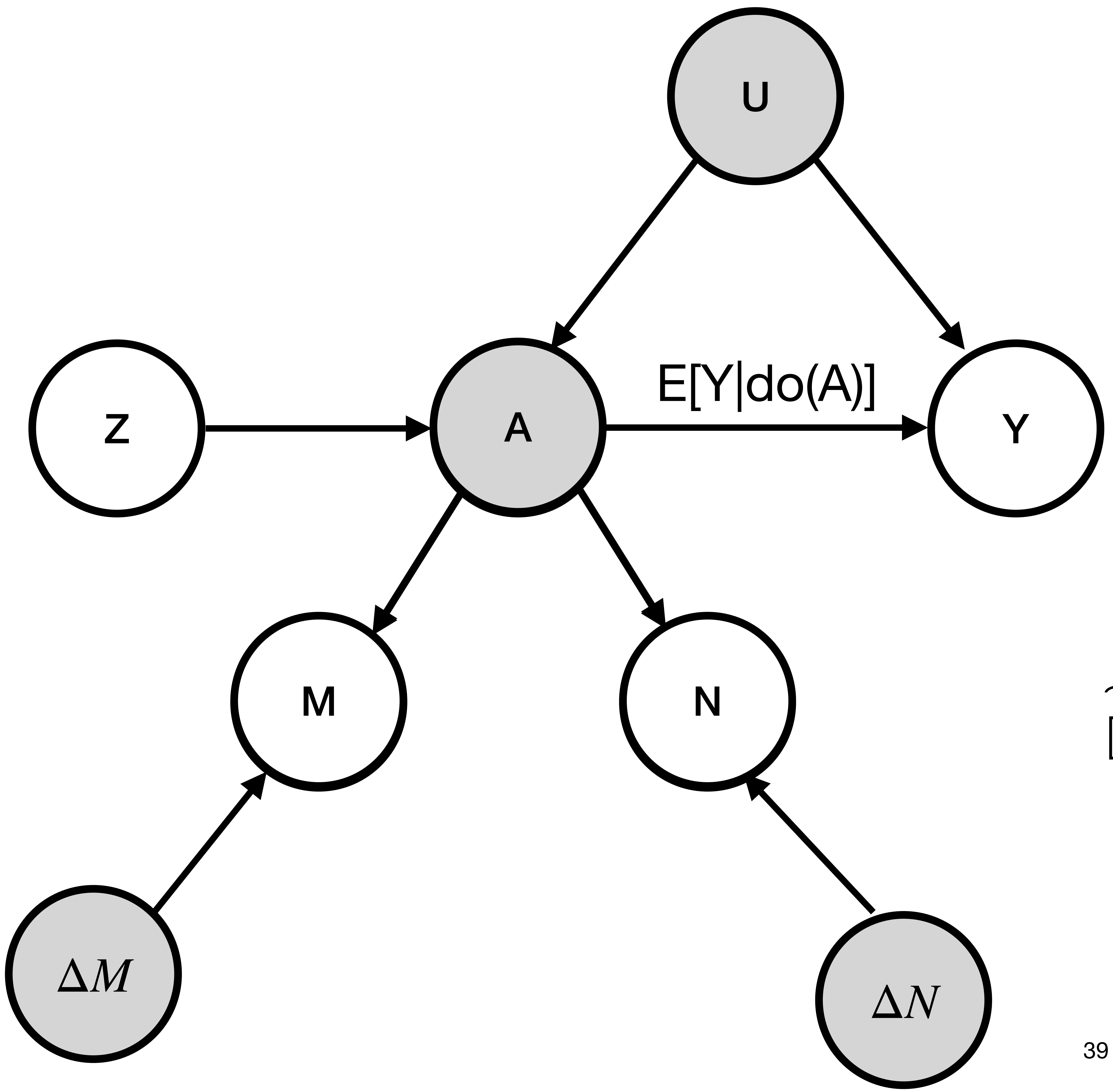
Application in causal inference with corrupted treatments



Application scenario

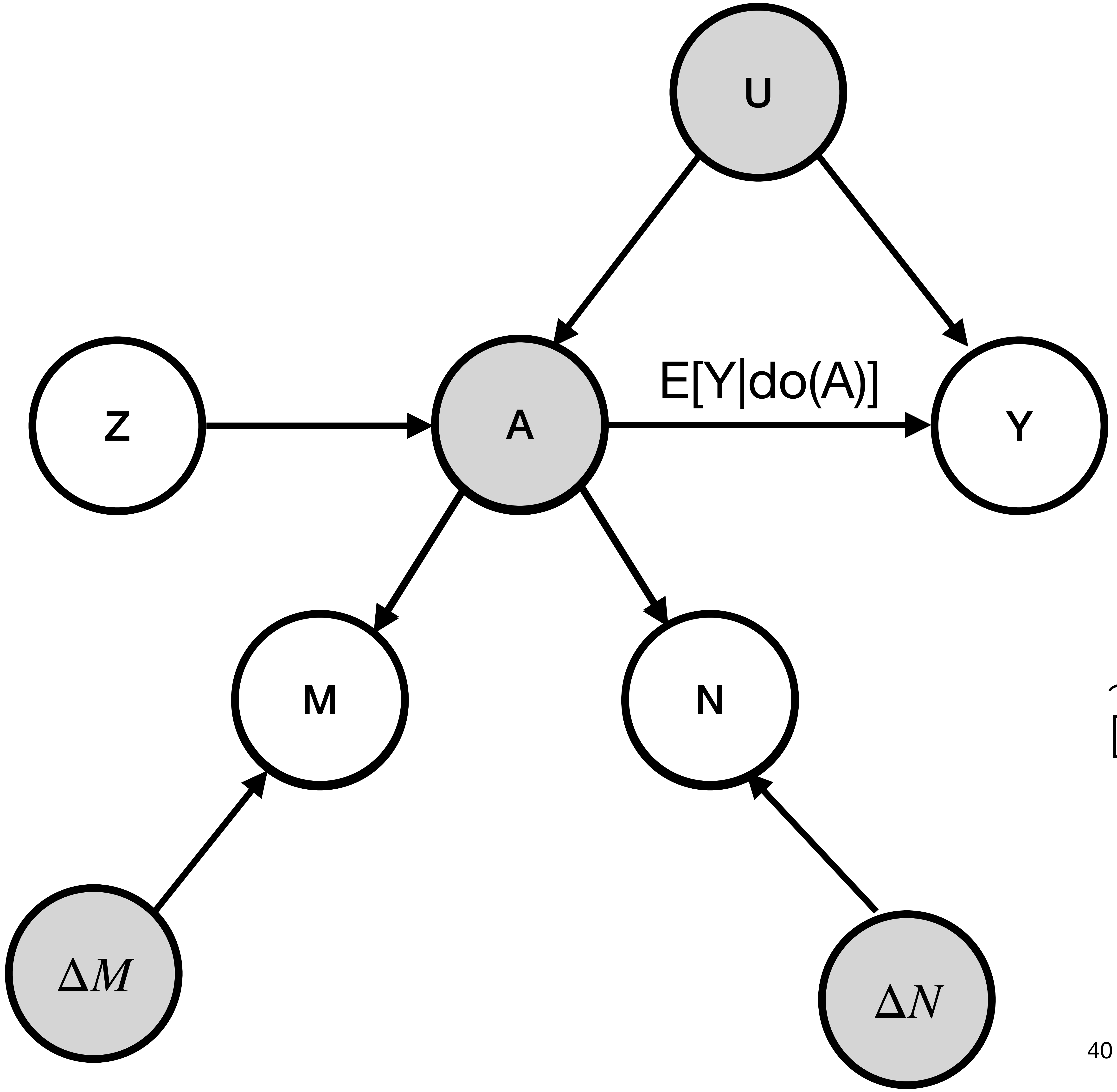


Application in causal inference with corrupted treatments



$$\overbrace{\mathbb{E}_{\mathscr{P}_{A|z}}\left[e^{i\alpha A}\,|\,z\right]}^{\psi_{\mathscr{P}_{A|z}}(\alpha):=} = \exp\left(\int_0^\alpha i\frac{\mathbb{E}\left[Me^{i\nu N}\,|\,z\right]}{\mathbb{E}\left[e^{i\nu N}\,|\,z\right]}d\nu\right)$$

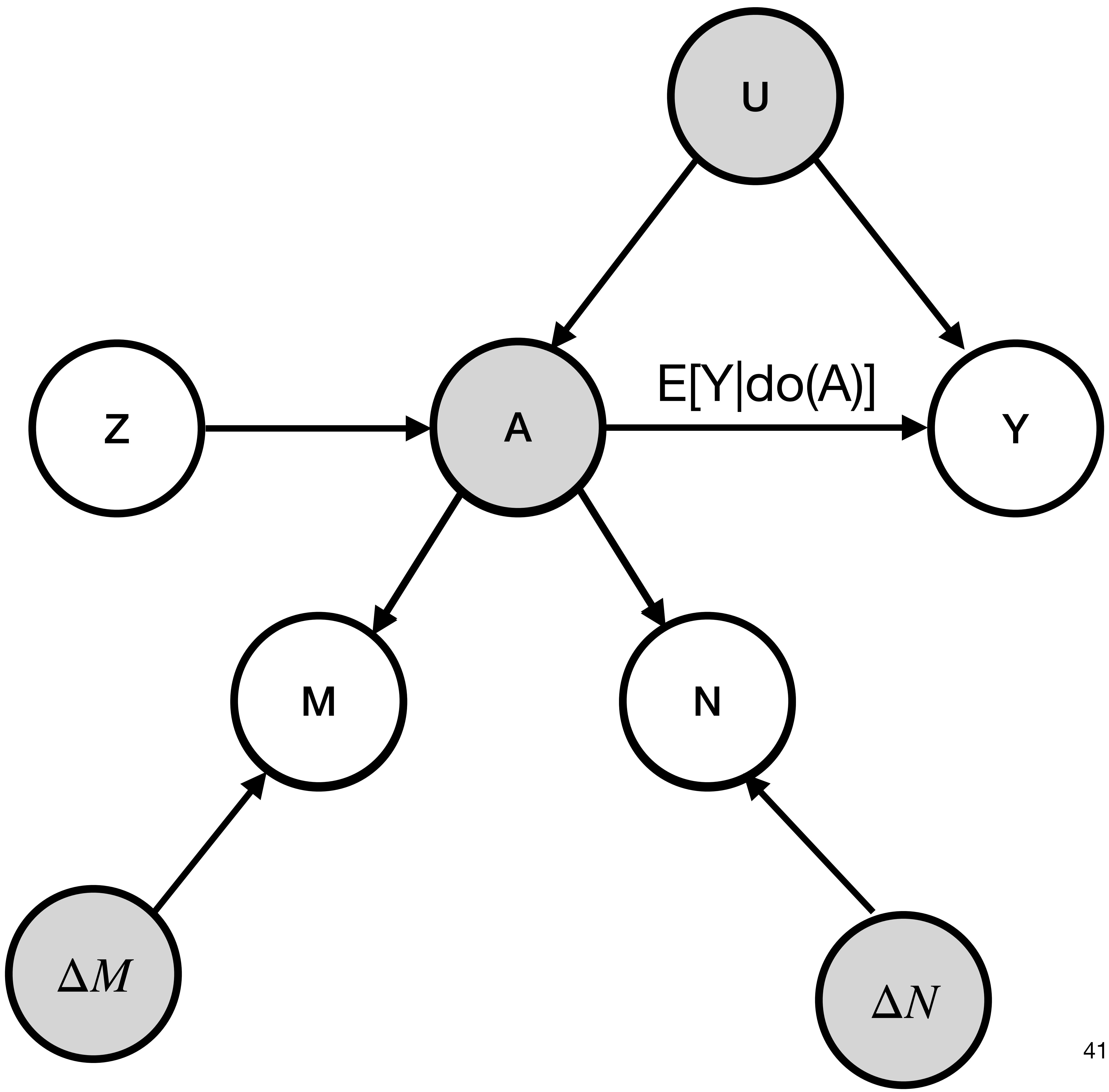
Application in causal inference with corrupted treatments



How to compute the right hand side?

$$\underbrace{\mathbb{E}_{\mathscr{P}_{A|z}}\left[e^{i\alpha A}\,|\,z\right]}^{\psi_{\mathscr{P}_{A|z}}(\alpha):=} = \exp\left(\int_0^\alpha i\frac{\mathbb{E}\left[Me^{i\nu N}\,|\,z\right]}{\mathbb{E}\left[e^{i\nu N}\,|\,z\right]}d\nu\right)$$

Application in causal inference with corrupted treatments



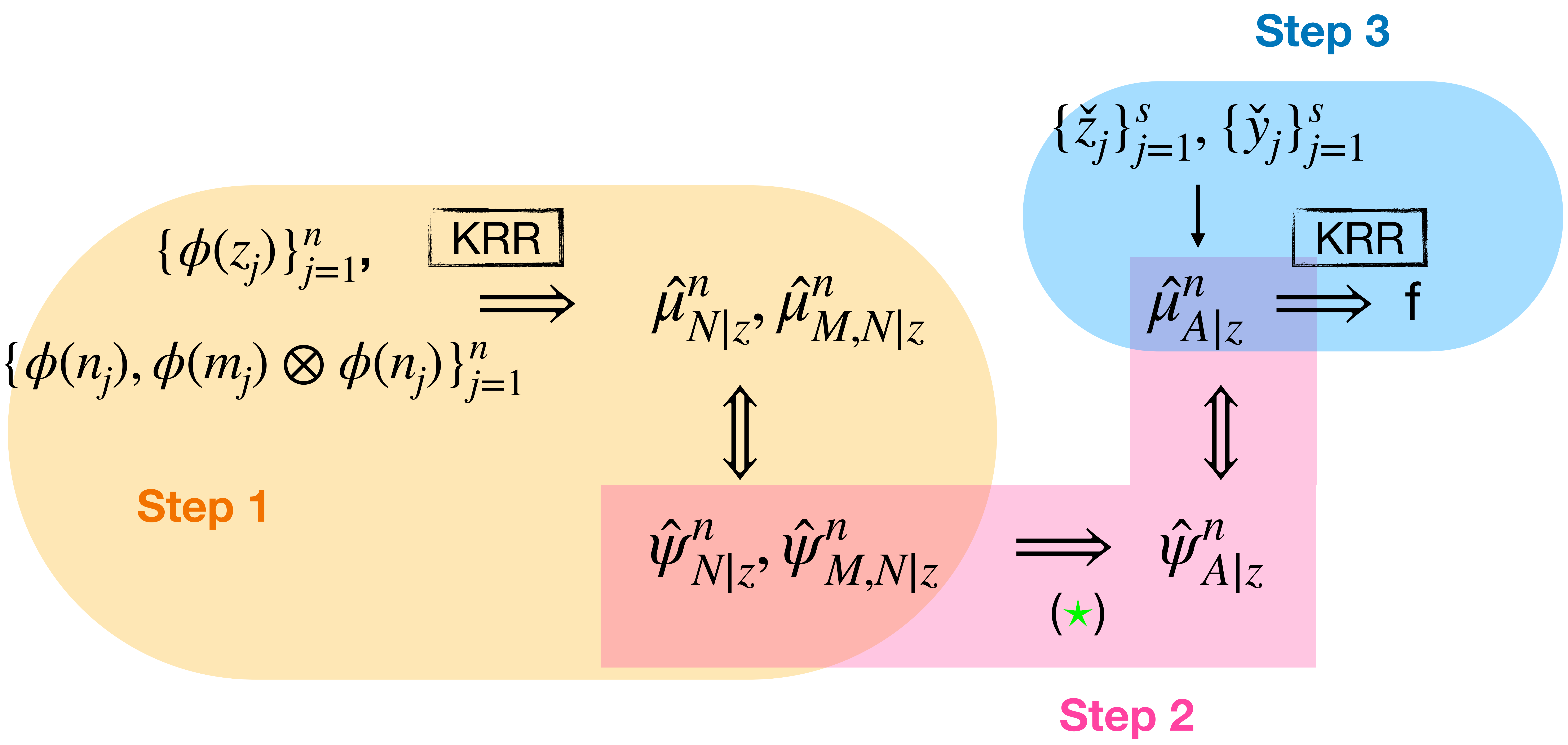
To obtain $\hat{\psi}^n_{A|z}$:

$$\overbrace{\mathbb{E}_{\mathscr{P}_{A|z}}[e^{i\alpha X}](\alpha)}^{\psi_{A|z}(\alpha)} = \exp\left(\int_0^\alpha i\frac{\overbrace{\mathbb{E}[Me^{i\nu N}\,|\,z]}^{\frac{\partial}{\partial v}\psi_{M,N|z}(v,\nu)\big|_{v=0}}}{\underbrace{\mathbb{E}[e^{i\nu N}\,|\,z]}_{\psi_{N|z}(\nu)}}d\nu\right) \quad (1)$$

1. **Differentiate wrt α to remove integral.**
2. **Replace with sample estimates.**

$$\frac{\frac{d}{d\alpha}\hat{\psi}^n_{A|z}(\alpha)}{\hat{\psi}^n_{A|z}(\alpha)} = \frac{\frac{\partial}{\partial v}\hat{\psi}^n_{M,N|z}(v,\alpha)\Big|_{v=0}}{\hat{\psi}^n_{N|z}(\alpha)} \quad (2)$$

Measurement Error KIV (MEKIV)

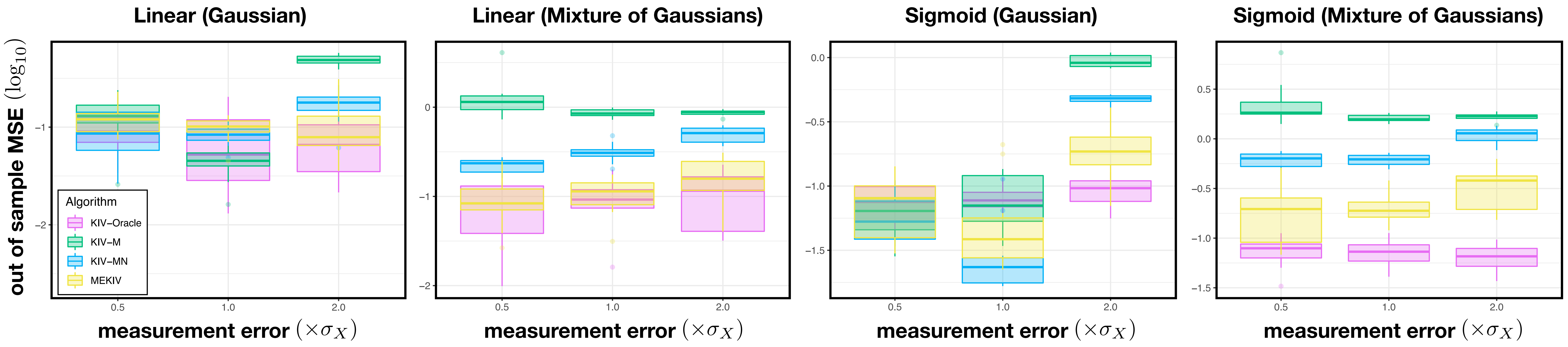


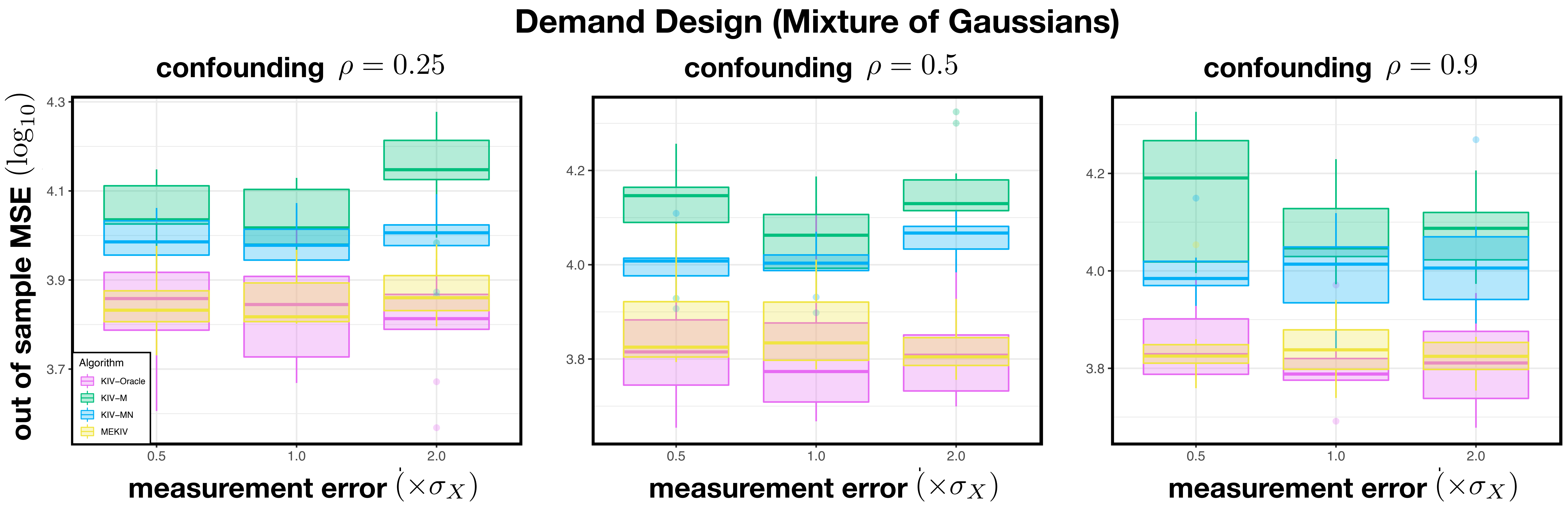
Zhu et al, UAI 2022, Causal Inference with Treatment Measurement Error: A Nonparametric IV Approach.

Advantages of MEKIV

- 无关于分布的假设 对Kotlarski假设的放宽: Evdokimov and White 2011.
- 极少的超参数调参。
- 对CME建模，而非对整个分布函数建模。

MEKIV results





Summary of techniques and future work

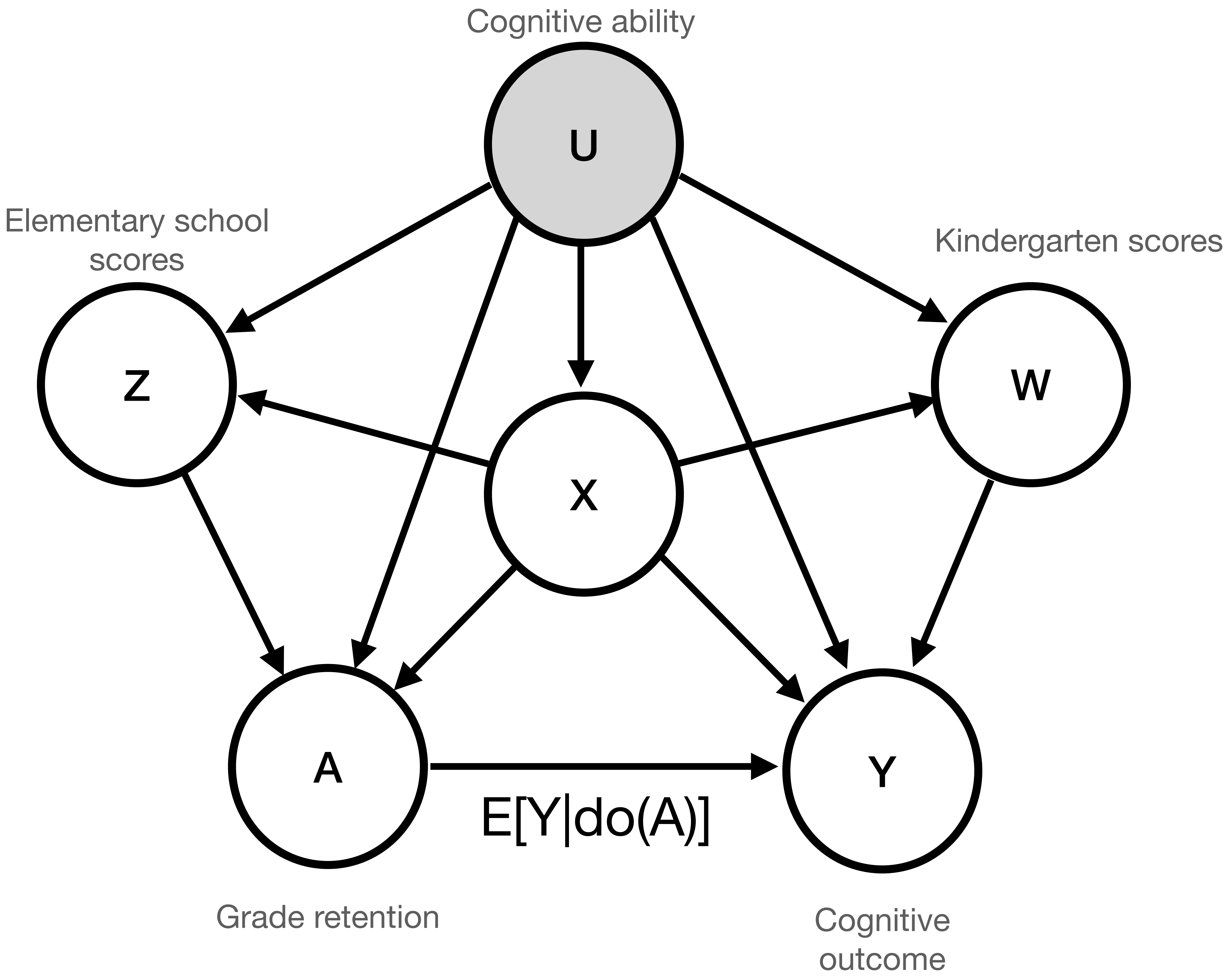
- Kotlarski引理允许我们从它们的两个线性组合中识别三个看不见的变量。这是否可以被继续探索?

- 特征函数和均值嵌入之间的对偶性是否可以带来更多两个方向的融合?

- 需要放宽加性误差假设。

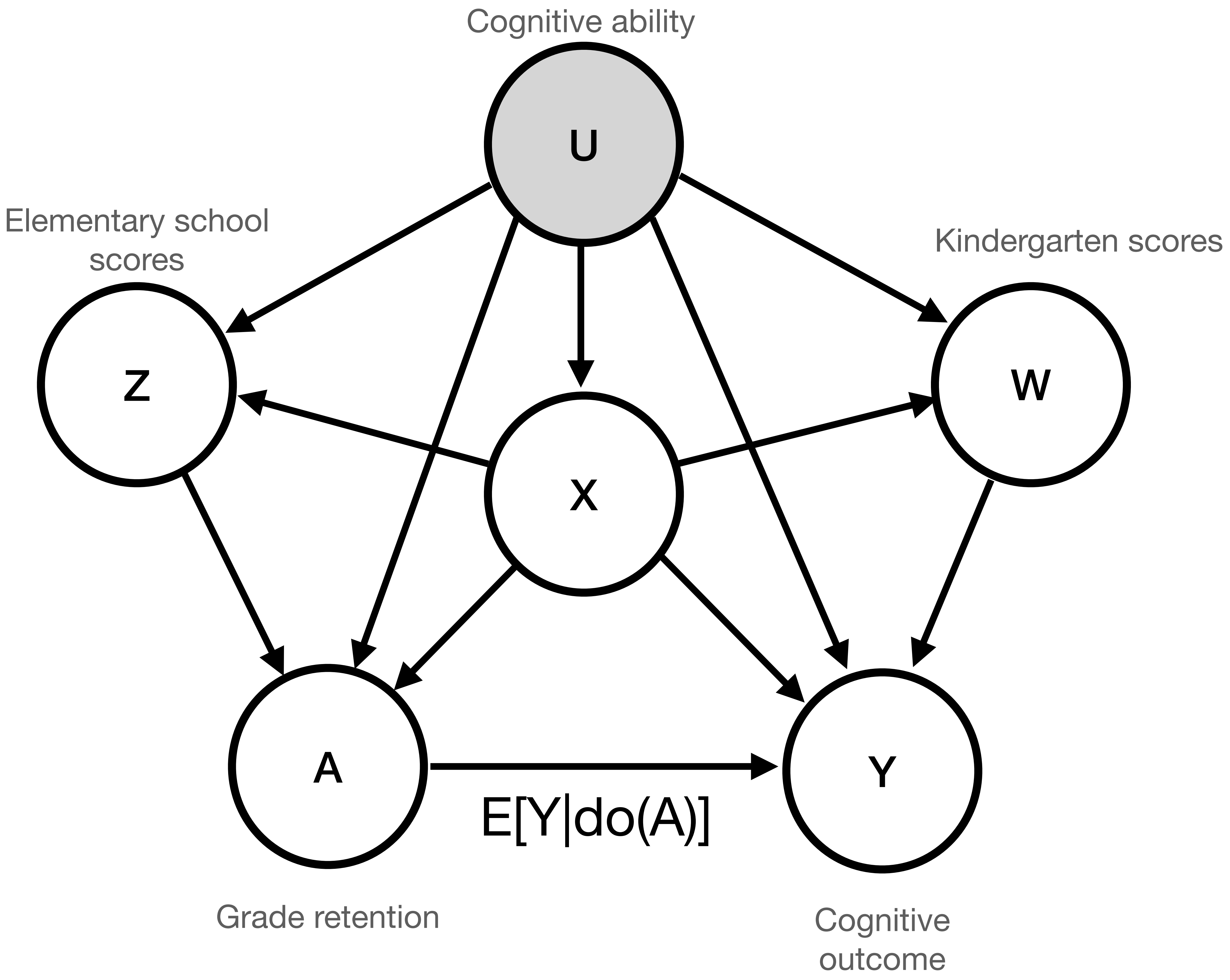
- 需要放宽instrumental variable假设。

Proximal Causal Learning Background



Tchetgen-Tchetgen et al 2020. An Introduction to Proximal Causal Learning.

Proximal Causal Learning Background

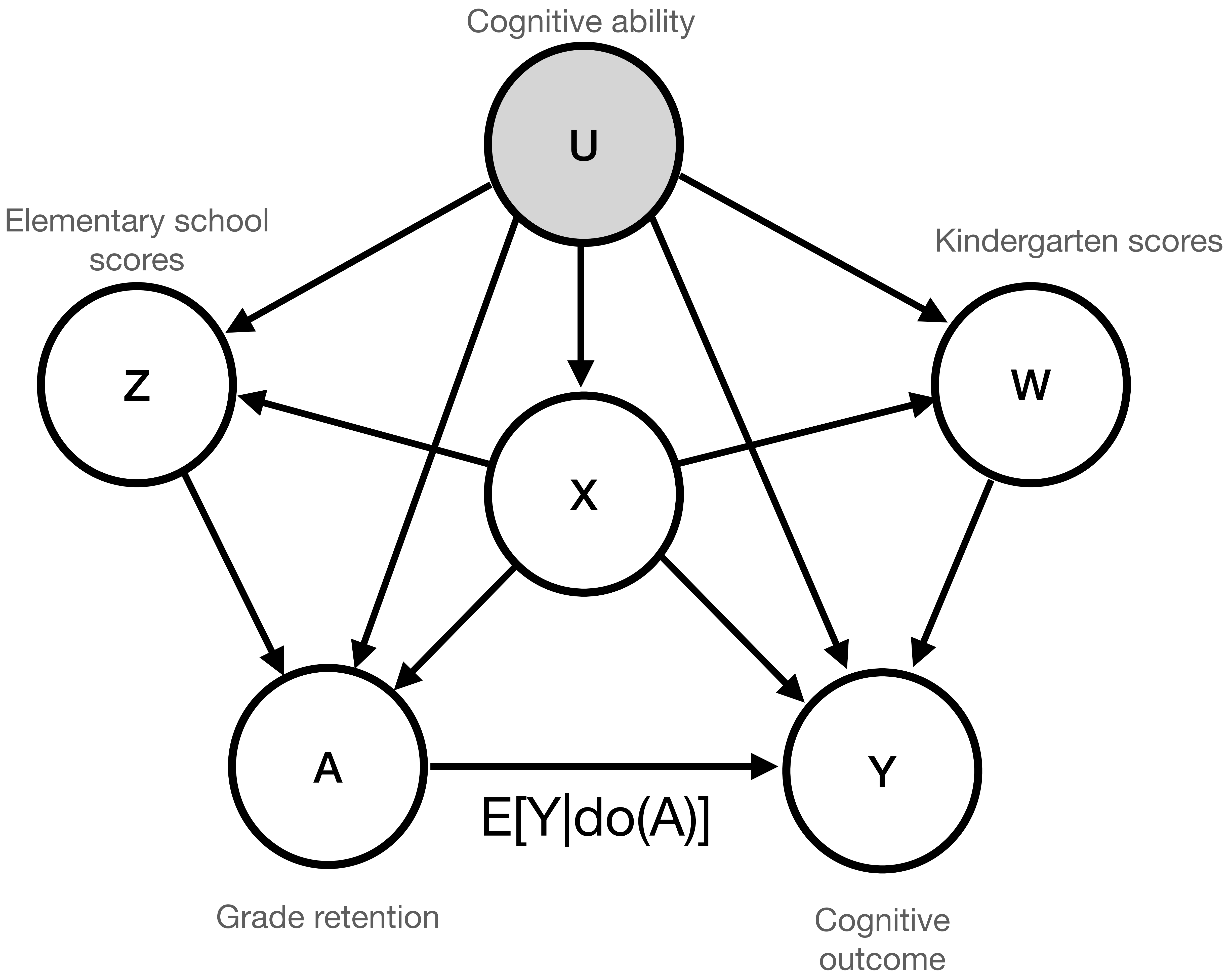


Average causal effect estimation:

$$\mathbb{E}[Y | do(A = a)] = \int_{XW} h(a, w, x)p(w, x)dxdw$$

Tchetgen-Tchetgen et al 2020. An Introduction to Proximal Causal Learning.

Proximal Causal Learning Background



Average causal effect estimation:

$$\mathbb{E}[Y\,|\,do(A=a)] = \int_{XW} h(a,w,x)p(w,x)dxdw$$

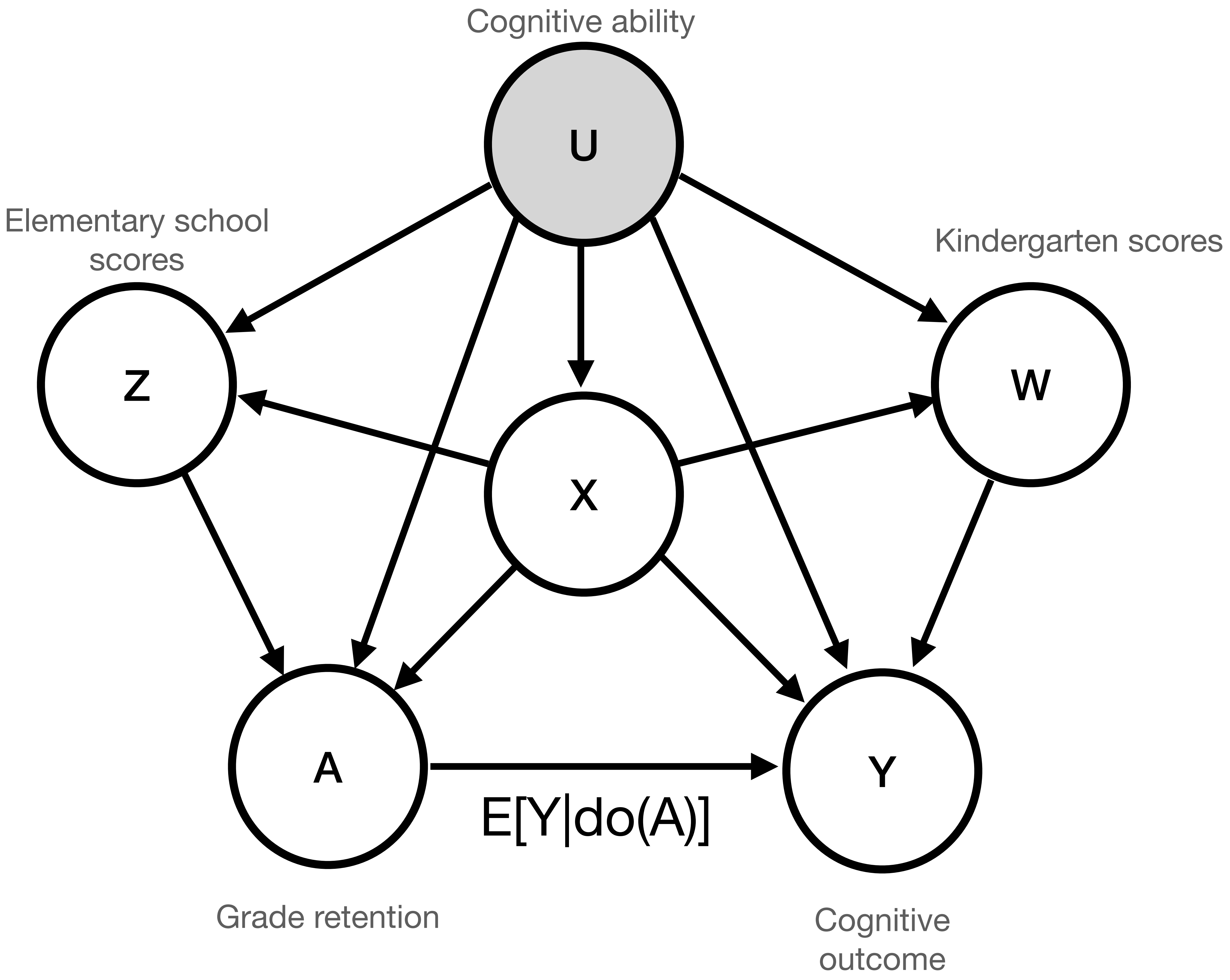
How to get h?

$$\Downarrow$$

$$\mathbb{E}[Y - h(A,W,X)\,|\,A,Z,X] = 0 \quad \text{a.s. } P_{AZX}$$

Tchetgen-Tchetgen et al 2020. An Introduction to Proximal Causal Learning.

Proximal Causal Learning Background



Average causal effect estimation:

$$\mathbb{E}[Y \mid do(A = a)] = \int_{XW} h(a, w, x) p(w, x) dx dw$$

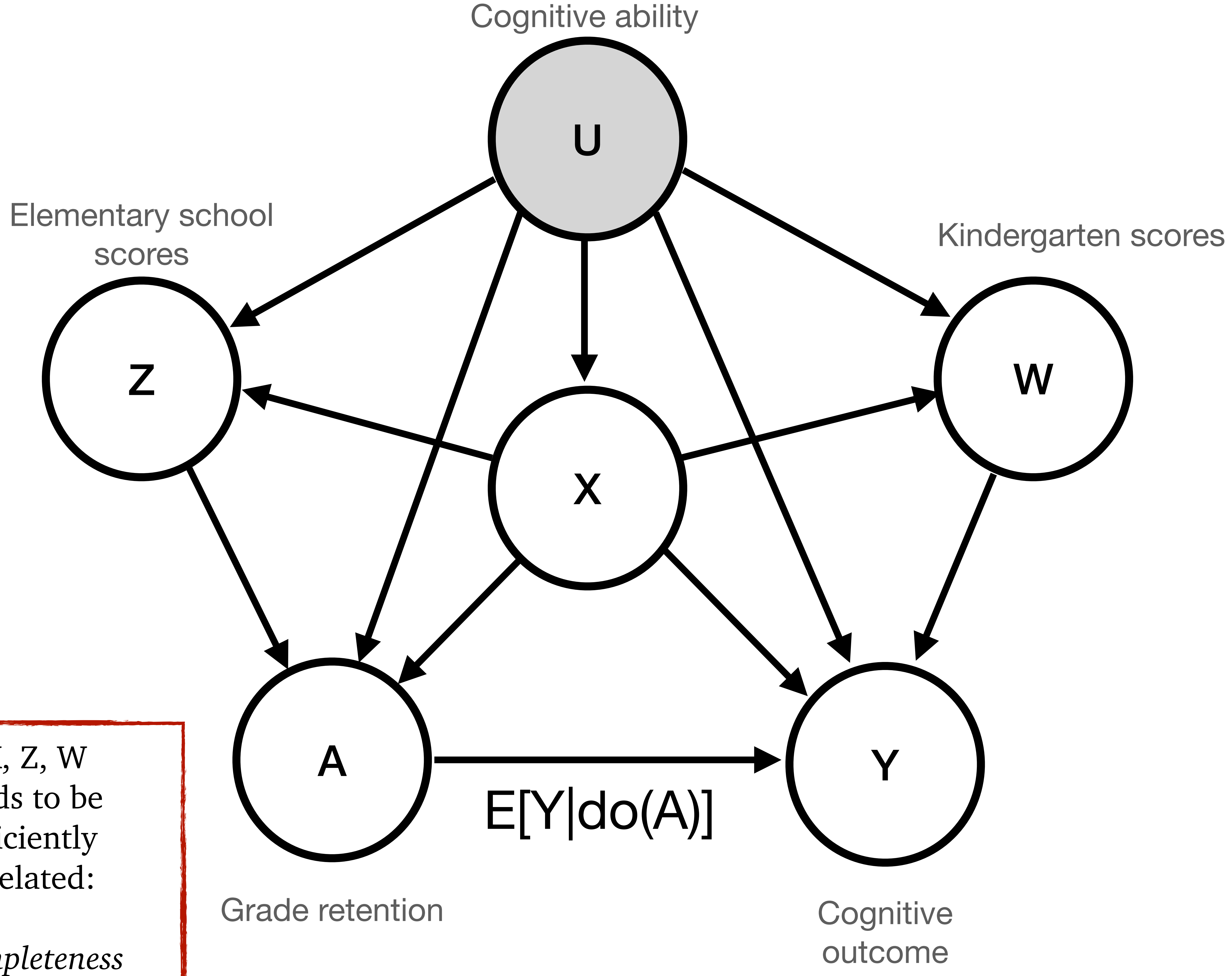
How to get h?

- **Expectation operator:** $\mathbb{E}[g(\cdot_U) \mid A, Z, X]$
- $\mathbb{E}[Y \mid A, U, X] = \int h(A, w, x) p(w, x \mid U, X) dx dw$

$\mathbb{E}[Y - h(A, W, X) \mid A, Z, X] = 0$ a.s. P_{AZX}

Tchetgen-Tchetgen et al 2020. An Introduction to Proximal Causal Learning.

Proximal Causal Learning Background



U, X, Z, W needs to be sufficiently correlated:

Completeness Condition (Miao et al. 2018)

Average causal effect estimation:

$$\mathbb{E}[Y \mid do(A=a)] = \int_{XW} h(a,w,x)p(w,x)dxdw$$

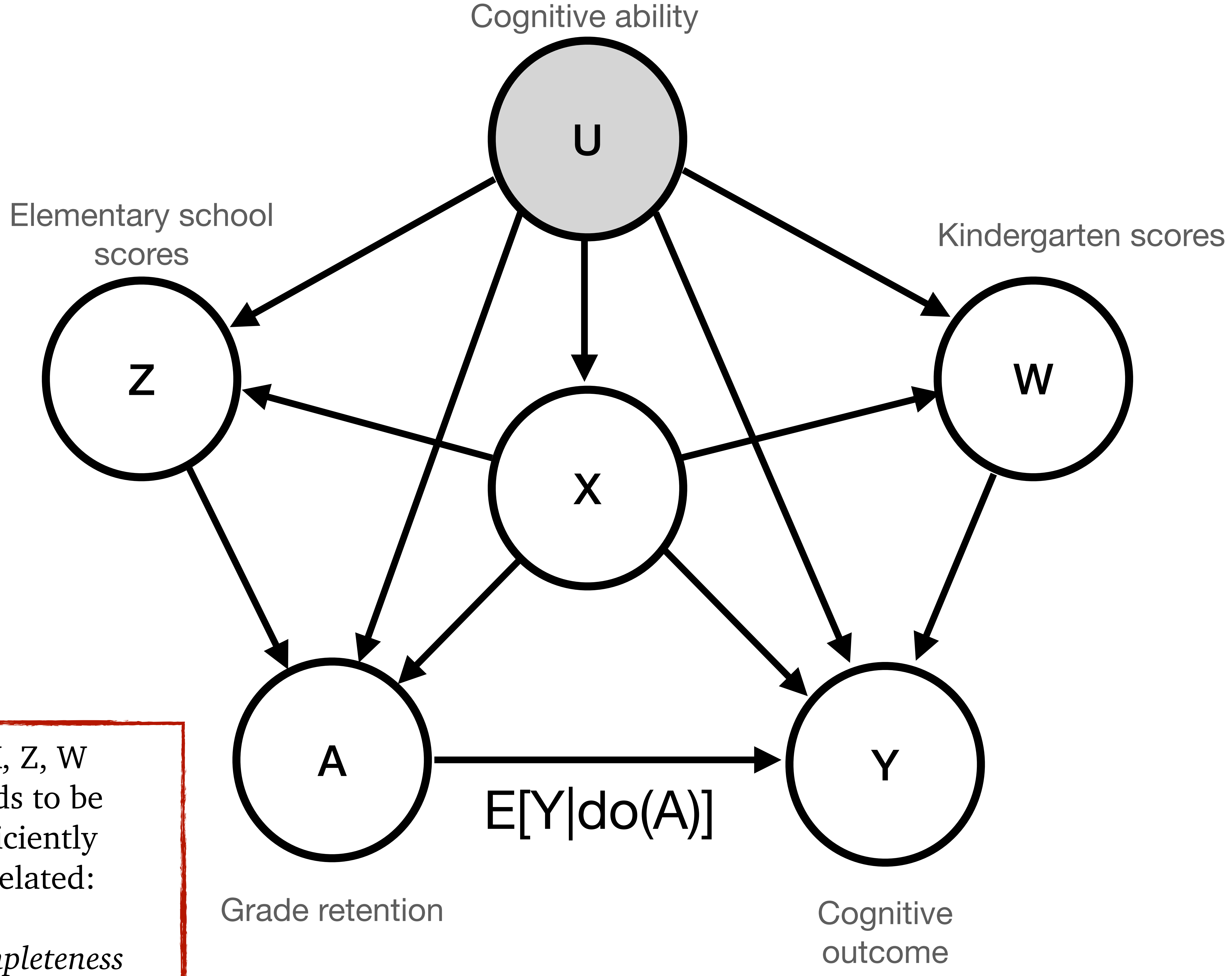
How to get h?

- **Expectation operator:** $\mathbb{E}[g(\cdot_U) \mid A, Z, X]$
- $\mathbb{E}[Y \mid A, U, X] = \int h(A, w, x)p(w, x \mid U, X)dxdw$

$$\mathbb{E}[Y - h(A, W, X) \mid A, Z, X] = 0 \quad \text{a.s. } P_{AZX}$$

Tchetgen-Tchetgen et al 2020. An Introduction to Proximal Causal Learning.

Proximal Causal Learning Background



U, X, Z, W needs to be sufficiently correlated:

Completeness Condition (Miao et al. 2018)

Average causal effect estimation:

$$\mathbb{E}[Y|do(A=a)] = \int_{XW} h(a,w,x)p(w,x)dxdw$$

How to get h?

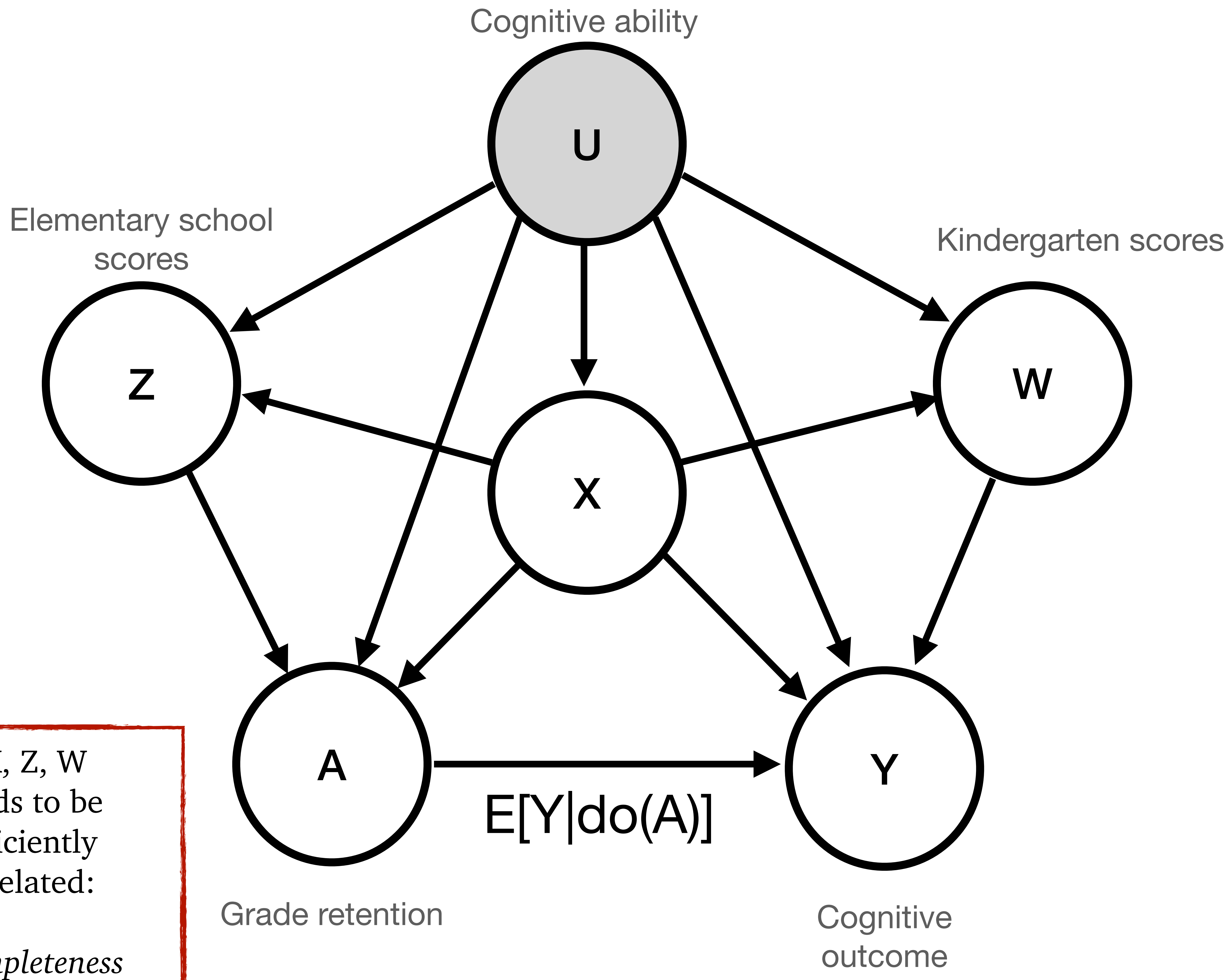
- Expectation operator: $\mathbb{E}[g(\cdot_U)\,|\,A,Z,X]$
- $\mathbb{E}[Y|A,U,X] = \int h(A,w,x)p(w,x\,|\,U,X)dxdw$

$$\mathbb{E}[Y - h(A,W,X)\,|\,A,Z,X] = 0 \quad \text{a.s. } P_{AZX}$$

- Normal regression equation: "$\mathbb{E}[Y - h(A,Z,X)\,|\,A,Z,X] = 0$ a.s. P_{AZX}"
- Here we also need to take the expectation over $P_{W|AZX}$.

Tchetgen-Tchetgen et al 2020. An Introduction to Proximal Causal Learning.

Proximal Causal Learning Background



U, X, Z, W needs to be sufficiently correlated:

Completeness Condition (Miao et al. 2018)

Average causal effect estimation:

$$\mathbb{E}[Y \mid do(A=a)] = \int_{XW} h(a,w,x)p(w,x)dxdw$$

How to get h?

- **Expectation operator:** $\mathbb{E}[g(\cdot_U) \mid A, Z, X]$
- $\mathbb{E}[Y \mid A, U, X] = \int h(A, w, x)p(w, x \mid U, X)dxdw$

$$\mathbb{E}[Y - h(A, W, X) \mid A, Z, X] = 0 \quad \text{a.s. } P_{AZX}$$

- **Normal regression equation:** "$\mathbb{E}[Y - h(A, Z, X) \mid A, Z, X] = 0$ a.s. P_{AZX}"
- **Here we also need to take the expectation over $P_{W|AZX}$.**

Tchetgen-Tchetgen et al 2020. An Introduction to Proximal Causal Learning.

Proximal Maximum Moment Restriction

$$\mathbb{E}[Y - h(A, X, W) \mid A, X, Z] = 0 \quad \text{a.s. } P_{AXZ}$$

CMR

$$\Updownarrow$$

$$\mathbb{E}[(Y - h(A, X, W))g(A, X, Z)] = 0 \quad \textbf{a.s. } P_{AXZ}$$ For all g

Mastouri*, **Z.***, et al. Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restrictions. *ICML 2021.*

Proximal Maximum Moment Restriction

$$\mathbb{E}[Y - h(A, X, W) \mid A, X, Z] = 0 \quad \text{a.s. } P_{AXZ}$$

CMR

$$\Updownarrow$$

- If E[A|B] = 0,
- Then (for g measurable):
- E[Ag(B)] = E[E[Ag(B)|B]]
- = E[E[A|B]g(B)] = 0

$$\mathbb{E}[(Y - h(A, X, W))g(A, X, Z)] = 0 \ \ \textbf{a.s. } P_{AXZ}$$

For all g

Mastouri*, **Z.***, et al. Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restrictions. *ICML 2021.*

Proximal Maximum Moment Restriction

$$\mathbb{E}[Y - h(A, X, W) \mid A, X, Z] = 0 \quad \text{a.s. } P_{AXZ}$$

CMR

$$\Updownarrow$$

- If E[A|B] = 0,
- Then (for g measurable):
- E[Ag(B)] = E[E[Ag(B)|B]]
- = E[E[A|B]g(B)] = 0

$$\mathbb{E}[(Y - h(A, X, W))g(A, X, Z)] = 0 \quad \textbf{a.s. } P_{AXZ}$$ For all g

Precursor loss:

$$R(h) = \sup_g (\mathbb{E}[(Y - h(A, W, X))g(A, Z, X)])^2$$

Mastouri*, **Z.***, et al. Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restrictions. *ICML 2021.*

Proximal Maximum Moment Restriction

$$\mathbb{E}[Y - h(A, X, W) \mid A, X, Z] = 0 \quad \text{a.s. } P_{AXZ}$$

CMR

$$\Downarrow$$

- If E[A|B] = 0,
- Then (for g measurable):
- E[Ag(B)] = E[E[Ag(B)|B]]
- = E[E[A|B]g(B)] = 0

$$\mathbb{E}[(Y - h(A, X, W))g(A, X, Z)] = 0 \quad \textbf{a.s. } P_{AXZ}$$

For all g

Precursor loss:

$$R(h) = \sup_g (\mathbb{E}[(Y - h(A, W, X))g(A, Z, X)])^2$$

$$\Updownarrow$$

- Restrict g to $\mathscr{H}_{\mathscr{A}\mathscr{X}\mathscr{Z}}$

PMMR surrogate loss $R_k(h)$ k indexes the kernel.

Mastouri*, **Z.***, et al. Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restrictions. *ICML 2021.*

Proximal Maximum Moment Restriction

Precursor loss:

$$R(h) = \sup_g (\mathbb{E}[(Y - h(A, W, X))g(A, Z, X)])^2$$

$$\Downarrow$$

- Restrict g to $\mathscr{H}_{\mathscr{A}\mathscr{X}\mathscr{Z}}$

$$R_k(h) = \sup_{g \in \mathscr{H}_{\mathscr{A}\mathscr{Z}\mathscr{X}}, \quad \|g\| \leq 1} (\mathbb{E}[(Y - h(A, W, X))\langle g, k((A, Z, X), \cdot)\rangle])^2$$

$$= \mathbb{E}[(Y - h(A, W, X))(Y' - h(A', W', X'))k((A, Z, X), (A', Z', X'))]$$

Proximal Maximum Moment Restriction

Precursor loss:

$$R(h) = \sup_g (\mathbb{E}[(Y - h(A, W, X))g(A, Z, X)])^2$$

$$\Downarrow$$

- Restrict g to $\mathscr{H}_{\mathscr{A}\mathscr{X}\mathscr{Z}}$

$$R_k(h) = \sup_{g \in \mathscr{H}_{\mathscr{A}\mathscr{Z}\mathscr{X}}, \quad \|g\| \leq 1} (\mathbb{E}[(Y - h(A, W, X))\langle g, k((A, Z, X), \cdot)\rangle])^2$$

$$= \mathbb{E}[(Y - h(A, W, X))(Y' - h(A', W', X'))k((A, Z, X), (A', Z', X'))]$$

V-statistic: $R_V(h) := \frac{1}{n^2} \sum_{i,j=1}^{n} (y_i - h_i)(y_j - h_j)k_{ij}$ (reweighed ERM!)

Mastouri*, **Z.***, et al. Proximal Causal Learning with Kernels: Two-Stage Estimation and Moment Restrictions. *ICML 2021.*