Non-Parametric Structure Learning on Hidden Tree-Shaped Distributions

[1]Konstantinos E. Nikolakakis, [2]Dionysios S. Kalogerias,
[1]Anand D. Sarwate

[1]Department of Electrical & Computer Engineering, Rutgers University
[2]Department of Electrical & Systems Engineering, University of Pennsylvania

Why learn from noisy data?

Data acquisition devices or sensors introduce noise

Local differential privacy

Communication constrains and quantization error

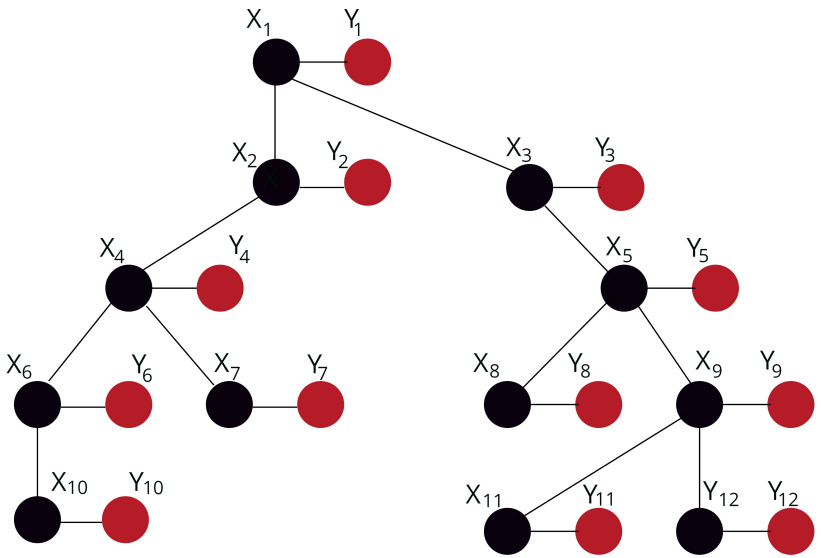
Adversarial attacks

Problem Statement (Learning Hidden Tree Structures)

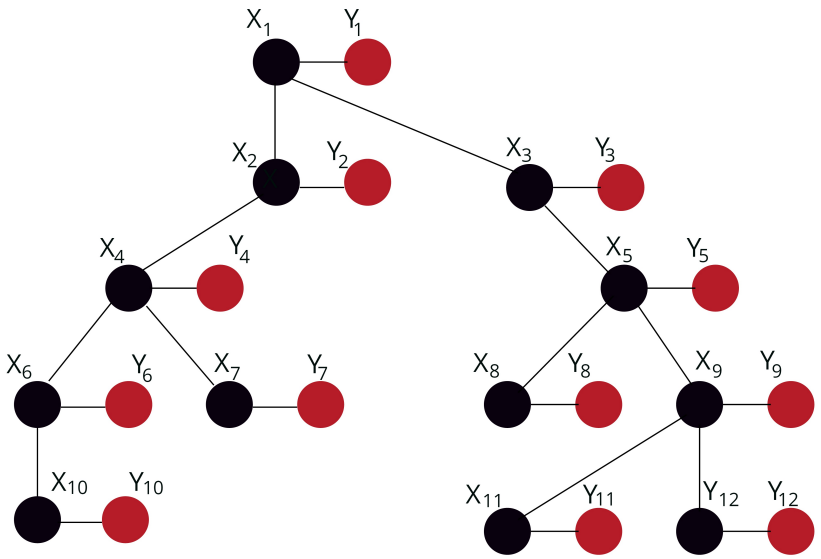
Observe noisy values for each node of the unknown tree structure

$X_1; X_2; : : : ; X_p$ are hidden variables (black nodes)

$Y_1; Y_2; : : : ; Y_p$ are observable variables (red nodes)



Learning a tree structure



Assumptions:

Distribution of X is nondegenerate and factorizes according to a tree T.

T = $(V; E)$ is connected.

$I(X_i; X_j) > 0$ for all $i; j \ 2 \ V$.

Chow-Liu Algorithm

Given: Data set $D = \boldsymbol{Y}\ 2\ Y^{jVj\ \ n}$

1. Compute empirical distribution on each edge:

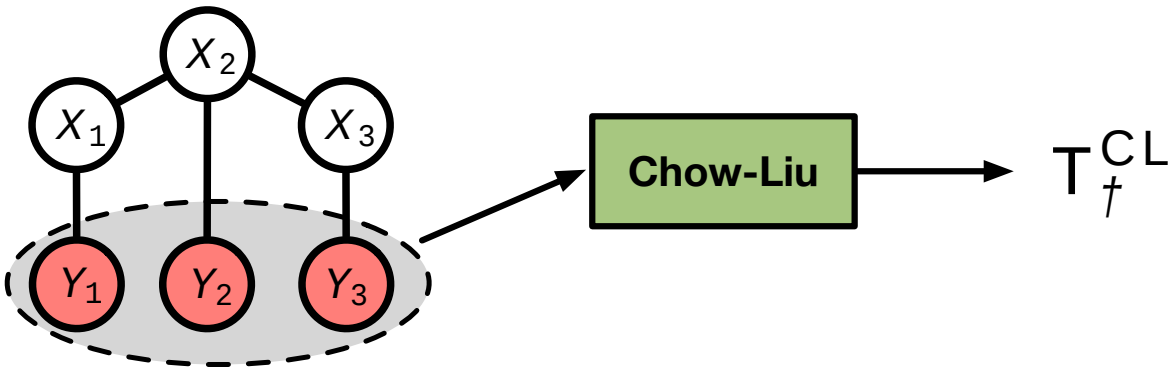
$$\hat{p}_{i;j}(`;m) = \frac{1}{n}\sum_{k=1}^{n} 1_{fY_{i;k}=`;Y_{j;k}=mg} \quad 8i;j\ 2\ V$$

2. Find plug-in estimate of mutual information:

$$\hat{I}(Y_i;Y_j) = \sum_{`;m} \hat{p}_{i;j}(`;m)\log_2 \frac{\hat{p}_{i;j}(`;m)}{\hat{p}_i(`)\hat{p}_j(m)}$$

3. Output $T_y^{\mathrm{CL}} = \mathrm{MST}\ \ f\hat{I}(Z_i;Z_j) : i;j\ 2\ Vg$

Main questions



Given noise corrupted data:

Is Chow-Liu consistent?

How does noise affect the sample complexity?

Prior work: Finite sample complexity for Ising and Gaussian Models.

Tan et al. (2011), Liu et al. (2011), Bresler & Karzand (2018)

Hidden models: Our work (2019), Goel-Kane-Klivans (2019)

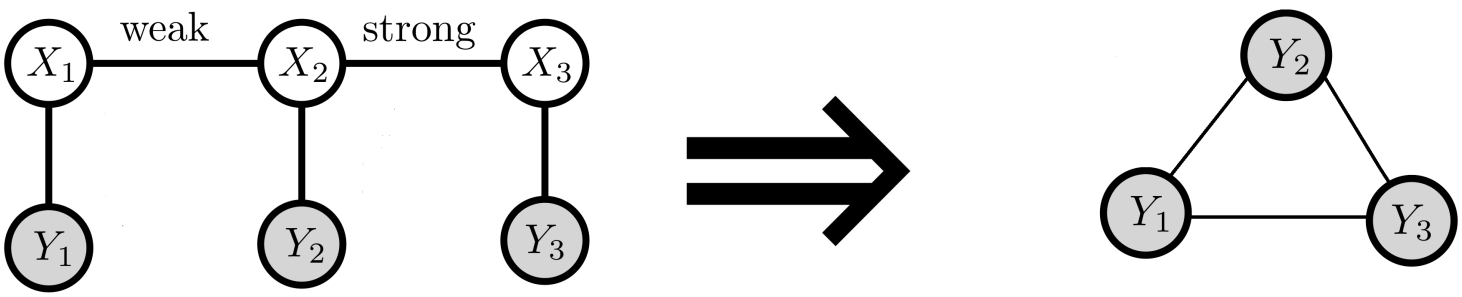
A motivating example: 3-node hidden model

What can go wrong when we have noise?

A motivating example: 3-node hidden model

What can go wrong when we have noise?

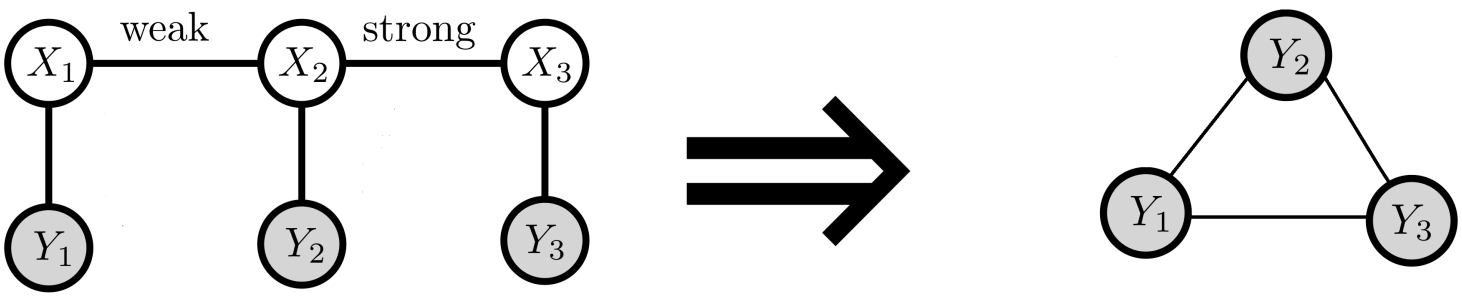
The MRF of the observable is a complete graph!



A motivating example: 3-node hidden model

What can go wrong when we have noise?

The MRF of the observable is a complete graph!

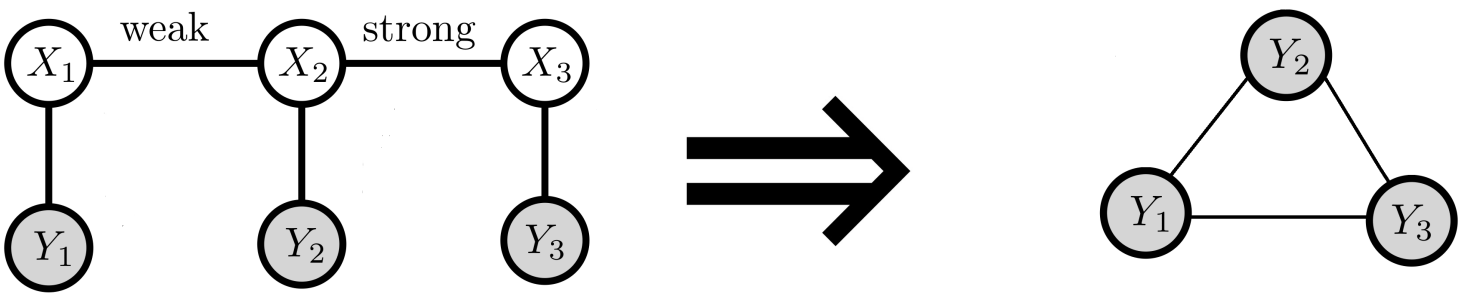


Questions:

A motivating example: 3-node hidden model

What can go wrong when we have noise?

The MRF of the observable is a complete graph!



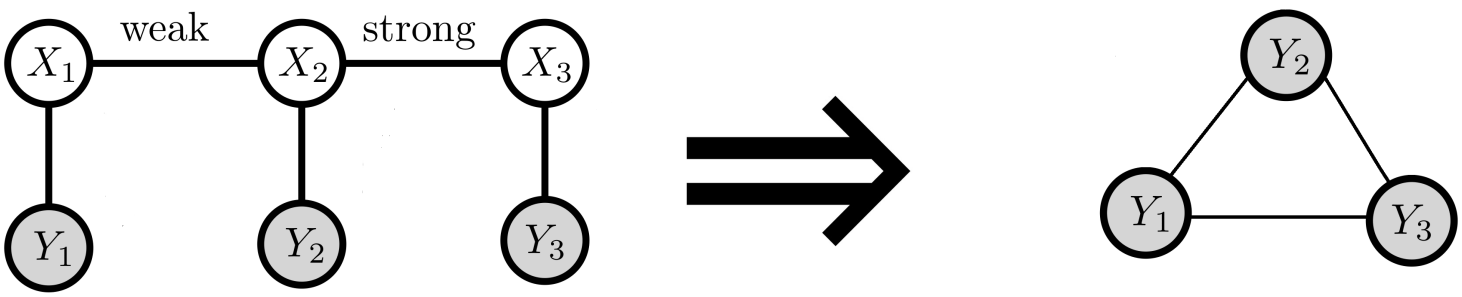
Questions:

Is Chow-Liu consistent? **NO**

A motivating example: 3-node hidden model

What can go wrong when we have noise?

The MRF of the observable is a complete graph!



Questions:

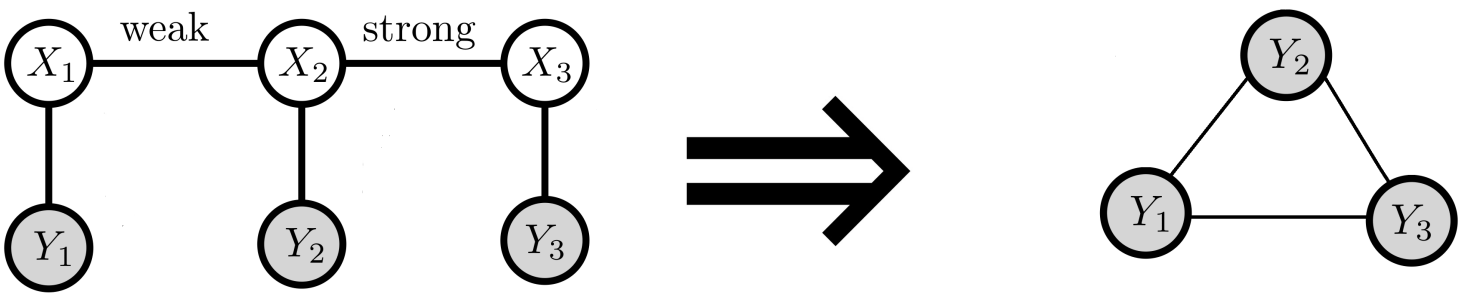
Is Chow-Liu consistent? **NO**

When does $\lim_{n \to \infty} T^{\mathrm{CL}} = T$ w.p. 1? **A su cient condition**

A motivating example: 3-node hidden model

What can go wrong when we have noise?

The MRF of the observable is a complete graph!



Questions:

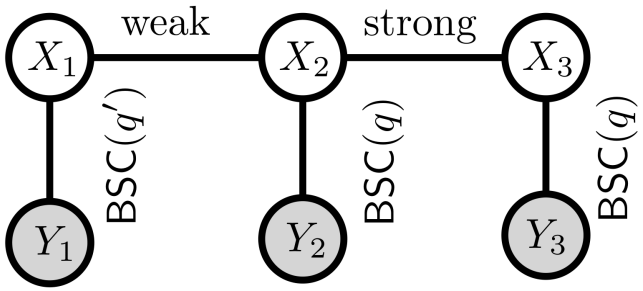
Is Chow-Liu consistent? **NO**

When does $\lim_{n \to \infty} T^{\mathrm{CL}} = T$ w.p. 1? **A su cient condition**

Can we tweak Chow-Liu to fix it? **Sometimes**

A closer look at the example

$$X_1, X_2, X_3 \in \{-1, +1\}, \quad 0 < \underbrace{|\mathrm{E}[X_1 X_2]|}_{\text{weak}} < \underbrace{|\mathrm{E}[X_2 X_3]|}_{\text{strong}} < 1$$



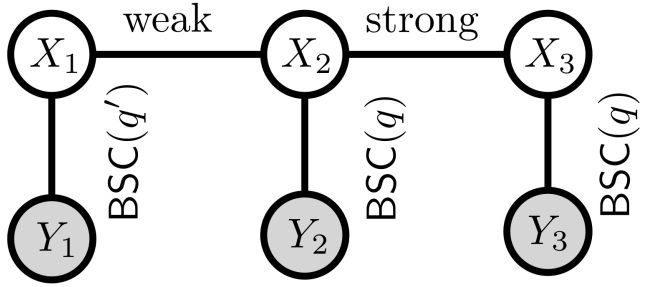
$$I(X_2; X_3) > I(X_1; X_2) \overset{\text{DPI}}{>} I(X_1; X_3)$$

$$\lim_{n \to \infty} \hat{I}(X_i; X_j) \to I(X_i; X_j) \text{ and } \lim_{n \to \infty} E_{\mathrm{T}^{\mathrm{CL}}} \to \{(1,2),(2,3)\} \quad E_{\mathrm{T}}$$

Does a similar condition hold for the observables?

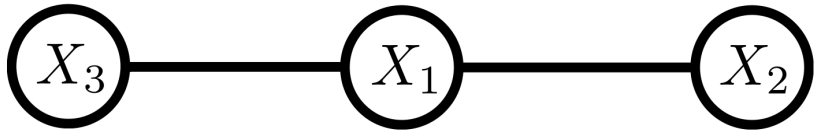
Could we have $\lim_{n \to \infty} E_{\mathrm{T}_y^{\mathrm{CL}}} \neq E_{\mathrm{T}}$?

Feasibility Threshold



If $I(Y_1;Y_2) > I(Y_1;Y_3) > I(Y_2;Y_3)$

then $\lim_{n \to \infty} E_{T_y^{CL}} \neq E_T$



$$I(Y_1;Y_3) > I(Y_2;Y_3) \quad ( ) \quad |E[Y_1Y_3]| > |E[Y_2Y_3]| \quad ( )$$

$$|E[X_1X_2]| > \frac{1 - 2q}{1 - 2q'}; \quad q; q' \in [0; 1/2):$$

Unprocessed vs Processed Data

What if $|\mathrm{E}[X_1 X_2]| > (1 \quad 2q)/(1 \quad 2q')$?

We have to pre-process

$$Z_1 \ , \ Y_1=(1 \quad 2q'); \ Z_2 \ , \ Y_2=(1 \quad 2q); \ Z_3 \ , \ Y_3=(1 \quad 2q)$$

Correct order, $I(Z_2; Z_3) > I(Z_1; Z_2) > I(Z_1; Z_3)$

Then $\lim_{n \to \infty} E_{\mathrm{T}_y^{\mathrm{CL}}} = E_{\mathrm{T}}$ with probability 1.

Unprocessed vs Processed Data

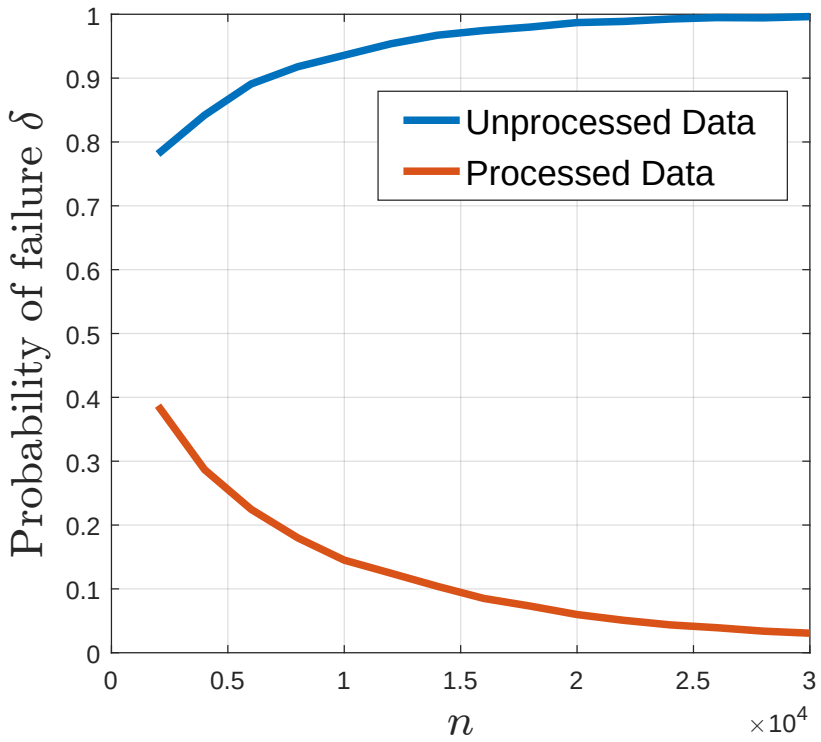
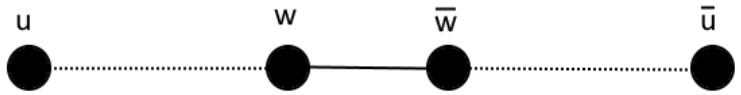


Figure: *Synthetic data, $q^0 = 0{:}2$, $q = 0{:}25$*

Definition

(**The set $\mathcal{EV}^2$**) Let $e \quad (w;w) \; 2 \; E_T$ be an edge and $u;u \; 2 \; V_T$ be a pair of nodes such that $e \; 2 \; \text{path}_T \, (u;u)$ and $j\text{path}_T \, (u;u) \, j \quad 2$. Then

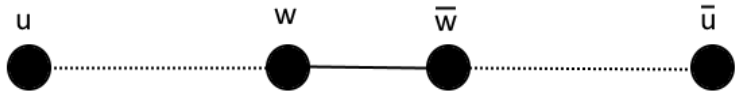
$$EV^2 \;\; , \;\; f(w;w);u;u \; 2 \; E_T \quad V_T \quad V_T :$$
$$(w;w) \; 2 \; \text{path}_T(u;u) \text{ and } j\text{path}_T \, (u;u) \, j \quad 2g:$$



Definition

(**The set $\mathcal{EV}^2$**) Let $e \quad (w;w) \in E_T$ be an edge and $u;u \in V_T$ be a pair of nodes such that $e \in \text{path}_T(u;u)$ and $|\text{path}_T(u;u)| \quad 2$. Then

$$EV^2 \ , \ f(w;w);u;u \in E_T \quad V_T \quad V_T :$$
$$(w;w) \in \text{path}_T(u;u) \text{ and } |\text{path}_T(u;u)| \quad 2g:$$



Error Characterization of CL algorithm (Bresler & Karzand 2018):

If T_y^{CL} ~~=~~ $T \quad =) \quad 9((w;w);u;u) \in EV^2 : \hat{I}(Y_w;Y_{\bar{w}}) \quad \hat{I}(Y_u;Y_{\bar{u}})$

Sufficient Condition for Exact Recovery

Exact recovery:

$$\text{If } \forall ((w,\bar{w}),(u,\bar{u})) \in E^{V^2} : \hat{I}(Y_w, Y_{\bar{w}}) > \hat{I}(Y_u, Y_{\bar{u}}) \implies T_y^{\mathrm{CL}} = T$$

$$\hat{I}(Y_w, Y_{\bar{w}}) > \hat{I}(Y_u, Y_{\bar{u}}) \quad (\ )$$

$$I(Y_w, Y_{\bar{w}}) - I(Y_u, Y_{\bar{u}}) > \left[\hat{I}(Y_u, Y_{\bar{u}}) - I(Y_u, Y_{\bar{u}})\right] - \left[\hat{I}(Y_w, Y_{\bar{w}}) - I(Y_w, Y_{\bar{w}})\right] :$$

Sufficient Condition for Exact Recovery

Exact recovery:

$$8((w;w);u;u)\ 2\ EV^2 : \hat{I}(Y_w;Y_{\bar{w}}) > \hat{I}(Y_u;Y_{\bar{u}}) \quad ( ) \quad \mathrm{T}_y^{\mathrm{CL}} = T$$

$$\hat{I}(Y_w;Y_{\bar{w}}) > \hat{I}(Y_u;Y_{\bar{u}}) \quad ( )$$

$$I(Y_w;Y_{\bar{w}}) \quad I(Y_u;Y_{\bar{u}}) > \left[\hat{I}(Y_u;Y_{\bar{u}}) \quad I(Y_u;Y_{\bar{u}})\right] \quad \left[\hat{I}(Y_w;Y_{\bar{w}}) \quad I(Y_w;Y_{\bar{w}})\right] :$$

Sufficient Condition

If $\hat{I}(Y_{\grave{}};Y_{\tilde{}}) \quad I(Y_{\grave{}};Y_{\tilde{}}) < \frac{1}{2} \min_{(e;u;\bar{u})2EV^2} \{I(Y_w;Y_{\bar{w}}) \quad I(Y_u;Y_{\bar{u}})\}$

for all $\grave{};\grave{}^{0}\ 2\ V$ then $\mathrm{T}_y^{\mathrm{CL}} = \mathrm{T}:$

Definition

(**Information Thresholds** I^o, I^o_y)

$$\mathsf{I}^o \ , \ \frac{1}{2} \min_{((w;\bar{w});u;\bar{u}) 2 EV^2} [I\,(X_w; X_{\bar{w}}) \quad I\,(X_u; X_{\bar{u}})]$$

$$\mathsf{I}^o_y \ , \ \frac{1}{2} \min_{((w;\bar{w});u;\bar{u}) 2 EV^2} [I\,(Y_w; Y_{\bar{w}}) \quad I\,(Y_u; Y_{\bar{u}})]$$

Always I^o 0, DPI

I^o_y 7 0 generalizes the condition $\frac{1 \quad 2q}{1 \quad 2q^0}$ 7 $j\mathsf{E}[X_1 X_2]j$ to non-parametric models and general channels

$\mathsf{I}^o_y < 0$ implies that structure learning is infeasible without post-processing

Sample Complexity

Sufficient condition $\hat{I}(Y_\ell; Y_{\ell'}) \quad I(Y_\ell; Y_{\ell'}) < I_y^o$

Concentration of measure of mutual information estimates

Union bound over the pairs $\ell, \ell' \in V$

Theorem

Fix $\in (0,1)$. *There exist constants* $C > 0$ *and* $c \in (1,2]$ *independent of* $$ *such that, if* $I_y^o > 0$ *and*

$$n \quad \frac{72 \log \frac{p}{}}{I_y^o \quad Cn^{\frac{1}{c}}}^2 \text{ and } I_y^o > Cn^{\frac{1}{c}};$$

then CL with input $D = Y^{1:n}$ *returns* $T_y^{CL} = T$ *w.p. at least* 1 .

Almost logarithmic order: $O \log^{1+}(p/)$, for all > 0

Experiments: Noiseless Binary Data

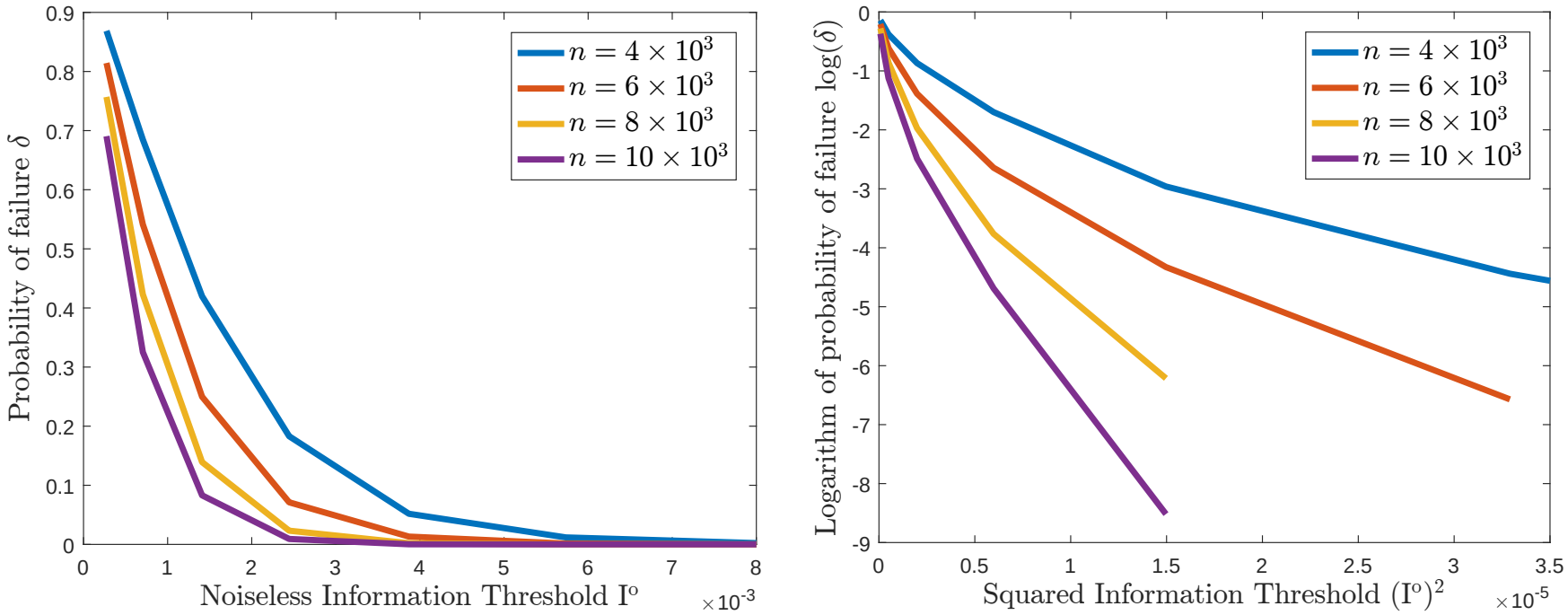


Figure: *Left:* $\hat{P}(T^{CL} \neq T)$ *vs* I^o, *Right:* $\log \hat{P}(T^{CL} \neq T)$ *vs* $(I^o)^2$

Experiments: Noisy Binary Data (BSC)

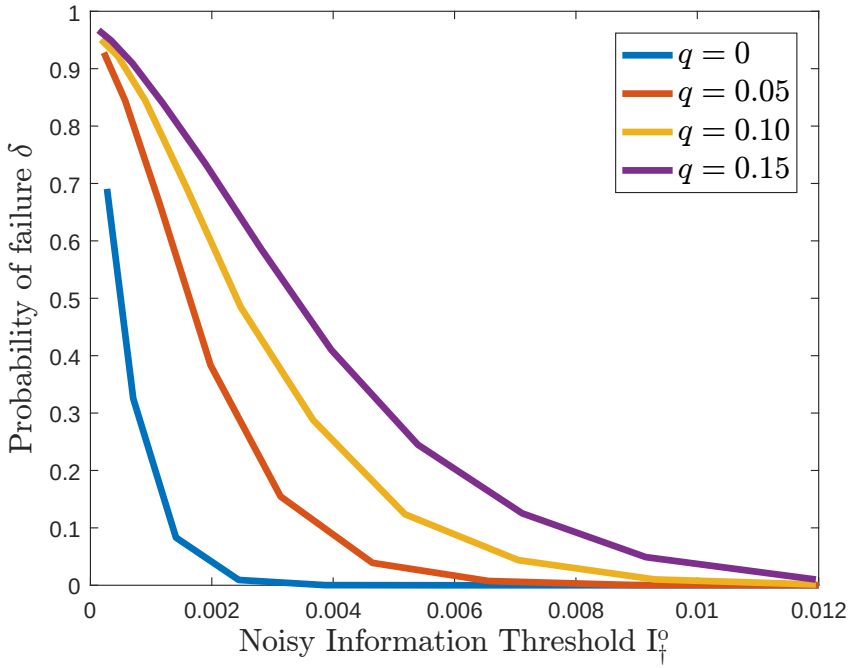


Figure: $\hat{P}$ T_y^{CL} ϵ T *vs* I_y^o

Further Questions and Future Directions

What is the relationship of I^o and I^o_y? Connection with SDPI

How to estimate I^o_y from training data?

How to preserve privacy while structure learning remains feasible?

Find robust methods for pre-processing against adversarial attacks

Thank you!