Valid Inference after Causal Discovery

Anonymous Author(s)
Affiliation
Address
email

Abstract

Causal graph discovery and causal effect estimation are two fundamental tasks in causal inference. While many methods have been developed for each task individually, statistical challenges arise when applying these methods jointly: estimating causal effects after running causal discovery algorithms on the same data leads to "double dipping," invalidating coverage guarantees of classical confidence intervals. To this end, we develop tools for valid post-causal-discovery inference. One key contribution is a randomized version of the greedy equivalence search (GES) algorithm, which permits a valid, distribution-free correction of classical confidence intervals. We show that a naive combination of causal discovery and subsequent inference algorithms typically leads to highly inflated miscoverage rates; at the same time, our noisy GES method provides reliable coverage control while achieving more accurate causal graph recovery than data splitting.

1 Introduction

Causal discovery and *causal estimation* are fundamental tasks in causal reasoning and decision-making. Causal discovery aims to identify the underlying structure of the causal problem, often in the form of a graphical representation which makes explicit which variables causally influence which other variables, while causal estimation aims to quantify the magnitude of the effect of one variable on another. These two goals frequently go hand in hand: computing valid causal effects requires adjustments that rely on either assuming or discovering the underlying graphical structure.

Methodologies for causal discovery and causal estimation have mostly been developed separately, and the statistical challenges that arise when solving these problems jointly have largely been overlooked. Indeed, a naive combination of causal discovery and standard methods for computing causal effects suffers from "double dipping": classical confidence intervals, such as those used for linear regression coefficients, need no longer cover the target estimand if the causal structure is not fixed a priori but is estimated on the same data used to compute the intervals. The key underlying problem is that *asserting the existence of a causal relationship biases the estimated effect size toward significance*.

More formally, suppose we are given a fixed causal graph G. Let β_G denote a causal parameter of interest within G, which will typically correspond to an effect of one variable on another. Standard statistical methods take a data set $\mathcal{D}$ and produce a confidence interval $\mathrm{CI}_G(\alpha; \mathcal{D})$ such that $\mathbb{P}\{\beta_G \notin \mathrm{CI}_G(\alpha; \mathcal{D})\} \leq \alpha$, where $\alpha \in (0, 1)$ is a pre-specified error level. However, if we *estimate* the causal graph $\widehat{G}$ from $\mathcal{D}$, this guarantee breaks down; that is, there is *no* guarantee that $\mathbb{P}\big\{\beta_{\widehat{G}} \notin \mathrm{CI}_{\widehat{G}}(\alpha; \mathcal{D})\big\} \leq \alpha$. This issue arises due to the coupling between the estimand $\beta_{\widehat{G}}$ and the data used for inference, since $\widehat{G}$ implicitly depends on $\mathcal{D}$.

To address this failure of naive inference, we develop tools for valid statistical inference after causal discovery. We build on concepts introduced in the literature on adaptive data analysis [Dwork et al., 2015a,b] and develop causal discovery algorithms that allow computing downstream confidence intervals with rigorous coverage guarantees. Our key observation is that *randomizing* causal discovery

Submitted to the NeurIPS Causal ML for Real-World Impact Workshop (CML4Impact 2022). Do not distribute.

mitigates the bias due to data reuse. In particular, we show that, for a level $\tilde{\alpha} \leq \alpha$ depending on the level of randomization, naive intervals satisfy $\mathbb{P}\{\beta_{\widehat{G}} \notin \mathrm{CI}_{\widehat{G}}(\tilde{\alpha}; \mathcal{D})\} \leq \alpha$, where $\widehat{G}$ is a causal structure estimated via a noisy causal discovery algorithm.

Randomization leads to a quantifiable tradeoff between the quality of the discovered structure and the statistical power of downstream inferences: higher levels of randomization imply lower structure quality, but at the same time allow tighter confidence intervals; that is, $\tilde{\alpha}$ is not much smaller than the target error level α. Moreover, we show empirically that the proposed randomization schemes are not vacuous: classical confidence intervals for causal effects indeed vastly undercover the target causal effect when computed after running standard, noiseless discovery algorithms.

A key contribution of our work is NOISY-GES, a noisy version of the classical *greedy equivalence search* (GES) [Chickering, 2002]. We show that NOISY-GES inherits consistency of usual GES, but at the same time enables a valid correction of classical confidence intervals in the learned graph.

2 Problem Formulation and Preliminaries

Causal Graphs. We consider the problem of performing inference based on a causal graph. A *causal* graph is a directed acyclic graph (DAG) $G = (V, E)$, where $V = (X_1, \ldots, X_d)$ is the set of vertices and E is the set of edges. We denote by $\mathbf{Pa}_j^G \subseteq [d]$ the set of parents of node X_j in graph G. In addition to capturing conditional independence relationships, a causal graph represents the causal relations in the data: the existence of an edge from X_i to X_j implies a possible causal effect from X_i to X_j. Our theory also applies to methods that return an *equivalence class* of DAGs, namely a *completed partially directed acyclic graph* (CPDAG). We will use the notation G, as well as the term "causal graph," to refer to both DAGs and CPDAGs, given that our tools are largely agnostic to whether the causal discovery criterion is applied to a set of possible DAGs or CPDAGs.

Targets of Inference in Causal Graphs. Suppose that the analyst works with a causal graph G and decides on a specific causal estimator to compute the effect of X_i on X_j within this graph. We will use $\beta_G^{(i \to j)}$ to denote the large-sample limit of this estimator, and that will be our target of inference in graph G. Analogously, when $\widehat{G}$ is discovered from data, our target will be $\beta_{\widehat{G}}^{(i \to j)}$.

It is natural ask whether inference—and specifically its resulting target—is meaningful if the discovered graph is not the exact generating truth, since then $\beta_{\widehat{G}}^{(i \to j)}$ may not coincide with the "true" causal effect. The perspective we build upon is that different models provide different *approximations* to the truth, some better than others, and should not be thought of as true data-generating processes [Berk et al., 2013, Buja et al., 2019a,b]. Indeed, a causal graph is rarely a perfect representation of the truth, but it can nevertheless serve as a useful working model. For instance, given the complexity of any real-world system, some relevant factors will almost inevitably be missing from the graph used in the analysis. This is true when the graph is estimated algorithmically, but even when it is provided by a domain expert. Whether or not the graph is correct, there is a well-defined underlying population-level quantity that the estimator approximates. Naturally, if the discovered graph $\widehat{G}$ is correct, then $\beta_{\widehat{G}}^{(i \to j)}$ will be equivalent to the true causal effect. The goal of our confidence interval constructions is to appropriately measure the estimator's fluctuations around the target.

Statistical Validity. To perform a causal analysis, we work with a finite data set $\mathcal{D} = \{X^{(k)}\}_{k=1}^n \equiv \{(X_1^{(k)}, \ldots, X_d^{(k)})\}_{k=1}^n$ of n i.i.d. measurements from a distribution $\mathcal{P}$, where $X_j^{(k)}$ denotes the j-th variable in data point k. With only finite data, valid inference is ensured by constructing *confidence intervals* around an estimator, often by relying on the estimator's (asymptotic) normality. See Imbens [2004] for an overview of standard confidence interval constructions. We study settings in which the causal graph G is not given a priori but is learned from $\mathcal{D}$ via causal discovery algorithms. Denote by $\widehat{G}$ the graph over $X_1, \ldots, X_d$ obtained in a data-driven way. Our main technical result can be summarized as follows: whenever we have a way of constructing valid confidence intervals for a causal quantity of interest when the causal graph G is *fixed*, we can adapt the respective method to produce valid confidence intervals when the causal graph $\widehat{G}$ is *learned from data*.

What makes inferring the targets $\beta_{\widehat{G}}^{(i \to j)}$ statistically challenging is the fact that we are using the data twice: once to estimate the causal structure $\widehat{G}$ and another time to compute a causal estimate

$\widehat{\beta}^{(i\to j)}_{\widehat{G}}$. This double-dipping phenomenon creates a bias: $\widehat{\beta}^{(i\to j)}_{\widehat{G}}$ can be far further from $\beta^{(i\to j)}_{\widehat{G}}$ than predicted by classical statistical theory. To correct this bias, we rely on quantifying the error increase of "naive" confidence intervals due to double dipping. In particular, consider a family of confidence intervals $\mathrm{CI}^{(i\to j)}_G(\alpha;\mathcal{D})$ that satisfies

$$\mathbb{P}\Big\{\exists(i,j)\in\mathcal{I}_G : \beta^{(i\to j)}_G \notin \mathrm{CI}^{(i\to j)}_G(\alpha;\mathcal{D})\Big\}\leq\alpha, \tag{1}$$

for all G and $\alpha\in(0,1)$. Importantly, note that, since G is fixed, the target estimand is trivially independent of the data $\mathcal{D}$. The guarantee (1) does *not* hold when $\widehat{G}$ is estimated from $\mathcal{D}$. Throughout the paper we will use $\mathrm{CI}^{(i\to j)}_{\widehat{G}}(\alpha)\equiv\mathrm{CI}^{(i\to j)}_{\widehat{G}}(\alpha;\mathcal{D})$ to denote "standard" intervals, which, if $\mathcal{D}$ is independent of $\widehat{G}$, satisfy the high-probability guarantee of Eq. (1).

Correcting Inferences via Max-Information. We show that naive intervals at a corrected level $\tilde{\alpha}\leq\alpha$ have error at most α. This construction is intrinsically tied to the degree of dependence between the data $\mathcal{D}$ and the learned graph $\widehat{G}$, as formalized via *max-information*.

Definition 1 (Max-information [Dwork et al., 2015a])**.** *Given $\gamma\in(0,1)$, the γ-approximate max-information between $\mathcal{D}$ and $\widehat{G}$ is $I^{\gamma}_{\infty}(\widehat{G};\mathcal{D}):=\max_{\mathcal{O}}\log(\mathbb{P}\Big\{(\widehat{G},\mathcal{D})\in\mathcal{O}\Big\}-\gamma)/\mathbb{P}\Big\{(\widehat{G},\tilde{\mathcal{D}})\in\mathcal{O}\Big\}$, where $\tilde{\mathcal{D}}$ is an i.i.d. copy of $\mathcal{D}$ and $\mathcal{O}$ is maximized over all measurable sets.*

A bound on $I^{\gamma}_{\infty}(\widehat{G};\mathcal{D})$ provides a way of bounding the probability of miscoverage when $\widehat{G}$ is estimated from $\mathcal{D}$, as long as we can control the same notion of error in *fixed* graphs G. One approach for bounding max-information extensively studied in the literature on adaptive data analysis is to make the causal discovery procedure *differentially private* [Dwork et al., 2006]. Roughly speaking, differential privacy requires that the output of a statistical analysis be randomized in a way that makes it insensitive to the replacement of a single data point.

Definition 2 (Differential privacy [Dwork et al., 2006])**.** *A randomized algorithm $\mathcal{A}$ is ϵ-differentially private for some $\epsilon\geq 0$ if for any two fixed data sets $\mathcal{D}$ and $\mathcal{D}'$ differing in at most one entry and any measurable set $\mathcal{O}$, we have $\mathbb{P}\{\mathcal{A}(\mathcal{D})\in\mathcal{O}\}\leq e^{\epsilon}\mathbb{P}\{\mathcal{A}(\mathcal{D}')\in\mathcal{O}\}$, where the probabilities are taken over the randomness of the algorithm.*

To translate privacy into max-information, we apply the following key result.

Proposition 1 (Dwork et al. [2015a])**.** *Suppose that algorithm $\mathcal{A}$ is ϵ-differentially private, and fix any $\gamma\in(0,1)$. Then, we have $I^{\gamma}_{\infty}(\mathcal{A}(\mathcal{D});\mathcal{D})\leq\frac{n}{2}\epsilon^2+\epsilon\sqrt{n\log(2/\gamma)/2}$.*

Putting everything together, we see that it suffices to perform causal discovery in a differentially private manner in order to perform valid statistical inference downstream. We thus reduce the problem of valid inference after causal discovery to one of developing algorithms for private causal discovery.

3 Noisy Causal Discovery

Exact Search. Our first step is to study a simple setting in which the set of candidate graphs is small enough that we can exhaustively enumerate and individually score all of them. The following section extends our theory to the more realistic setting of large numbers of candidate graphs.

Suppose we have a candidate set $\mathcal{G}$ of causal graphs that captures our uncertainty about which data-generating model to choose. To select a graph from $\mathcal{G}$, we specify a score function, $S(G,\mathcal{D})$, which takes as input a graph G and data set $\mathcal{D}$, and we select the graph with the maximum score, $\widehat{G}_*=\arg\max_{G\in\mathcal{G}}S(G,\mathcal{D})$. The score function $S(G,\mathcal{D})$ is typically formulated as some measure of compatibility between G and the relationships suggested by the data $\mathcal{D}$, such as the Bayesian information criterion (BIC). Note that $\widehat{G}_*$ depends on the data $\mathcal{D}$ and is thus random.

To enable valid inference after graph selection, we rely on a randomized selection rule. Under this rule, a simple correction to the target error level α suffices for rigorous downstream inference. The key step in designing the randomized graph selection is to compute the maximum score in the uncertainty set $\mathcal{G}$ in a differentially private manner. To accomplish this, one needs to consider the *sensitivity* of the score. The amount of necessary randomization is directly proportional to the sensitivity.

Algorithm 1 Noisy causal discovery

input : data set $\mathcal{D}$, set of graphs $\mathcal{G}$, privacy parameter ϵ, score function S with sensitivity τ

output: causal graph $\widehat{G}$

For all $G \in \mathcal{G}$, sample $\xi_G \overset{\text{i.i.d.}}{\sim} \text{Lap}\left(\frac{2\tau}{\epsilon}\right)$

Set $\widehat{G} \leftarrow \arg\max_{G \in \mathcal{G}} S(G, \mathcal{D}) + \xi_G$

Return $\widehat{G}$

Definition 3 (Score sensitivity)**.** *A score function $S(G, \mathcal{D})$ is τ-sensitive if for any graph $G \in \mathcal{G}$ and any two fixed data sets $\mathcal{D}$ and $\mathcal{D}'$ differing in at most one entry, we have $|S(G, \mathcal{D}) - S(G, \mathcal{D}')| \leq \tau$.*

Roughly speaking, score sensitivity bounds the influence that any single data point can have on the choice of the best-scoring graph within the uncertainty set. We present our *noisy causal discovery* algorithm formally in Algorithm 1, and state its privacy guarantee in the following proposition.

Proposition 2. *Noisy causal graph discovery (Algorithm 1) is ϵ-differentially private.*

Combined with Proposition 1, Proposition 2 implies a correction in the form of a discounted error level for confidence interval construction that ensures valid inference on effects estimated from $\widehat{G}$.

Theorem 1. *Suppose $\widehat{G}$ is selected via Algorithm 1 and fix $\gamma \in (0, \alpha)$. Then, we have $\mathbb{P}\Big\{\exists (i,j) \in \mathcal{I}_{\widehat{G}} : \beta_{\widehat{G}}^{(i \to j)} \notin \text{CI}_{\widehat{G}}^{(i \to j)}(\tilde{\alpha})\Big\} \leq \alpha$, for $\tilde{\alpha} = (\alpha - \gamma)\exp\{-n\epsilon^2/2 - \epsilon\sqrt{n\log(2/\gamma)/2}\}$.*

Greedy Search. To enable valid statistical inference after causal discovery via GES [Chickering, 2002], we next develop a private variant of GES that relies on randomization. The GES algorithm requires the existence of a *local score*; that is, we can write the score of a graph as a sum of "subscores" obtained by regressing each variable X_i on its parents in G: $S(G, \mathcal{D}) = \sum_{i=1}^{d} s(X_i, \mathbf{Pa}_i^G, \mathcal{D})$. Standard scoring criteria, such as the Bayesian information criterion, satisfy this condition. As before, we define an appropriate notion of sensitivity.

Definition 4 (Local score sensitivity)**.** *A local score function s is τ-*sensitive *if $\forall i \in [d], I \subseteq [d]$ and any two data sets $\mathcal{D}$ and $\mathcal{D}'$ that differ in a single entry, we have $|s(X_i, X_I, \mathcal{D}) - s(X_i, X_I, \mathcal{D}')| \leq \tau$.*

Below we formally state the NOISY-GES algorithm along with its privacy guarantees. We stress that this procedure is equally valid for greedy search over CPDAGs and greedy search over DAGs. We use the notation $\Delta S^+(e, G, \mathcal{D}) := S(G \cup e, \mathcal{D}) - S(G, \mathcal{D})$ and $\Delta S^-(e, G, \mathcal{D}) := S(G \backslash e, \mathcal{D}) - S(G, \mathcal{D})$.

Proposition 3. *Noisy GES (Algorithm 2) is $(2\epsilon_{\text{thresh}} + 2E_{\max}\epsilon_{\text{score}})$-differentially private.*

With Proposition 3 in hand, we can now ensure valid statistical inference after causal discovery. We state an analogue of Theorem 1 for NOISY-GES which shows how to discount the target miscoverage level in order to preserve validity after graph discovery via greedy search.

Theorem 2. *Suppose that $\widehat{G}$ is selected via Algorithm 2 and fix $\gamma \in (0, \alpha)$. Then, we have $\mathbb{P}\Big\{\exists (i,j) \in \mathcal{I}_{\widehat{G}} : \beta_{\widehat{G}}^{(i \to j)} \notin \text{CI}_{\widehat{G}}^{(i \to j)}(\tilde{\alpha})\Big\} \leq \alpha$, for $\tilde{\alpha} = (\alpha - \gamma)\exp\left(-2n(\epsilon_{\text{thresh}} + E_{\max}\epsilon_{\text{score}})^2 - (\epsilon_{\text{thresh}} + E_{\max}\epsilon_{\text{score}})\sqrt{2n\log(1/\gamma)}\right)$.*

Consistency of NOISY-GES. Additionally, we show that NOISY-GES inherits consistency of the standard GES algorithm. In other words, employing randomization for valid downstream inference

Algorithm 2 Noisy greedy equivalence search

input : data $\mathcal{D}$, max. number of edges $E_{\max}$, score S with local sensitivity τ, parameters $\epsilon_{score}, \epsilon_{\text{thresh}}$

output: causal graph $\widehat{G}$

Initialize $\widehat{G}$ to be an empty graph

Run forward pass $\widehat{G} \leftarrow \text{GreedyPass}(\widehat{G}, \mathcal{D}, E_{\max}, S, \tau, \epsilon_{\text{score}}, \epsilon_{\text{thresh}}, +)$

Run backward pass $\widehat{G} \leftarrow \text{GreedyPass}(\widehat{G}, \mathcal{D}, E_{\max}, S, \tau, \epsilon_{\text{score}}, \epsilon_{\text{thresh}}, -)$

Return $\widehat{G}$

Algorithm 3 GreedyPass

input : initial graph $\widehat{G}_0$, data $\mathcal{D}$, max. number of edges $E_{\max}$, score S with local sensitivity τ, parameters $\epsilon_{\text{score}}, \epsilon_{\text{thresh}}$, pass indicator $\texttt{sgn} \in \{+, -\}$

output: estimated causal graph $\widehat{G}$

Initialize $\widehat{G} \leftarrow \widehat{G}_0$ and sample noisy threshold $\nu \sim \text{Lap}\left(\frac{4\tau}{\epsilon_{\text{thresh}}}\right)$

for $t = 1, 2, \ldots, E_{\max}$ **do**

  Construct set $\mathcal{E}_t^{\texttt{sgn}}$ of valid ($\texttt{sgn}$)-operators

  For all $e \in \mathcal{E}_t^{\texttt{sgn}}$, compute $\Delta S^{\texttt{sgn}}(e, \widehat{G}, \mathcal{D})$ and sample $\xi_{t,e} \overset{\text{i.i.d.}}{\sim} \text{Lap}\left(\frac{4\tau}{\epsilon_{\text{score}}}\right)$

  Set $e_t^* = \arg\max_{e \in \mathcal{E}_t^{\texttt{sgn}}} \Delta S^{\texttt{sgn}}(e, \widehat{G}, \mathcal{D}) + \xi_{t,e}$

  if $\Delta S^{sgn}(e_t^*, \widehat{G}, \mathcal{D}) + \eta_t > \nu$ *where* $\eta_t \sim \text{Lap}\left(\frac{8\tau}{\epsilon_{\text{thresh}}}\right)$ **then**

    Apply operator e_t^* to $\widehat{G}$

  else

    break

  end

end

Return $\widehat{G}$

incurs a negligible cost in large samples under suitable conditions. As for standard GES, the key condition for consistency is that an increase in score corresponds to an actual increase in the graph's ability to capture the underlying structure, formalized via *local consistency* [Chickering, 2002].

We make a minor assumption that $\mathcal{P}$ comes from an exponential family and that there exists a DAG $G_*(\mathcal{P})$ that is a *perfect map* of $\mathcal{P}$, meaning that every independence constraint in $\mathcal{P}$ is implied by the structure $G_*(\mathcal{P})$ and every independence implied by the structure $G_*(\mathcal{P})$ holds in $\mathcal{P}$. If there exists a perfect map of $\mathcal{P}$, we say that $\mathcal{P}$ is DAG-perfect.

Proposition 4 (Consistency of NOISY-GES)**.** *Denote by $\widehat{G}_{\text{GES}}$ the output of standard GES on $\mathcal{D}$. Moreover, suppose that the local score function is τ-sensitive. Assume $\frac{\tau}{\epsilon_{\text{score}}} = o(1)$, and $\frac{\tau}{\epsilon_{\text{thresh}}} = o(1)$. Further, assume that for all DAGs G and for all edges e, $\Delta S^{sgn}(e, G, \mathcal{D}) \to_p \Delta s_{e,G}^{sgn}$ and that $\Delta s_{e,G}^{sgn} \neq \Delta s_{e',G'}^{sgn}$ unless $e = e'$ and $G = G'$, for $\texttt{sgn} \in \{+, -\}$. Then, if $E_{\max} \geq |E(\widehat{G}_{\text{GES}})|$, $\lim_{n\to\infty} \mathbb{P}\Big\{\widehat{G} = \widehat{G}_{\text{GES}}\Big\} = 1$. If, in addition, $\mathcal{P}$ is DAG-perfect, and the scoring criterion is locally consistent, we have $\lim_{n\to\infty} \mathbb{P}\Big\{\widehat{G} = G_*(\mathcal{P})\Big\} = 1$, where G_* is a perfect map of $\mathcal{P}$.*

4 Empirical Studies

We compare the standard, non-noisy GES method with our noisy GES (Algorithm 2). We focus on multivariate Gaussian observations normalized to have unit variance. The most commonly used scoring criterion for GES is the *Bayesian information criterion* (BIC). It satisfies the conditions required to guarantee consistency of GES, but has unbounded sensitivity in general. To justify the conditions in Proposition 4, we we use *clipping* to guarantee a bounded local score sensitivity.

Definition 5 (Clipped BIC)**.** *The local clipped BIC score with clipping parameter C is defined as*

$$s_{BIC}^C(X_j, X_{\mathbf{Pa}_j^G}, \mathcal{D}) = -\min_\theta \frac{1}{n} \sum_{k=1}^n \min\{(X_j^{(k)} - \sum_{s \in \mathbf{Pa}_j^G} \theta_s X_s^{(k)})^2, C\} - \frac{|\mathbf{Pa}_j^G|}{n} \log n. \quad (2)$$

Proposition 5 (Clipped BIC properties)**.** *The clipped BIC score* (2) *satisfies: (i) $\frac{C}{n}$-sensitivity of the local score s_{BIC}^C; and (ii) local consistency, assuming $C = \omega(1)$.*

The two properties in Proposition 5 imply that the clipped BIC score can simultaneously achieve local consistency and τ-local sensitivity for any $\tau = \omega(\frac{1}{n})$. Therefore, to satisfy the conditions of Proposition 4 that ensure consistent graph recovery—in particular $\frac{\tau}{\epsilon_{\text{score}}}, \frac{\tau}{\epsilon_{\text{thresh}}} = o(1)$—we can use any $\epsilon_{\text{score}}, \epsilon_{\text{thresh}} = \omega(\frac{1}{n})$ and achieve consistency by calibrating C appropriately.

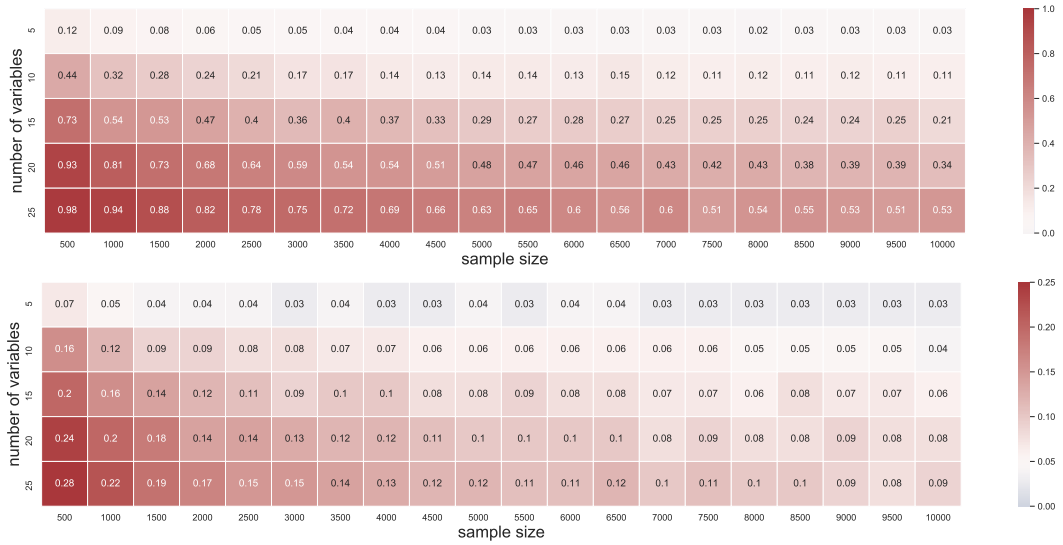

Figure 1. Probability of error for varying n (x-axis) and d (y-axis) in empty (top) and sparse random (bottom) graphs. Intervals are constructed with target error probability equal to 0.05. We observe that the probability of error significantly exceeds the target when the number of variables go beyond 10, even in large sample regimes.

Figure 2. Comparison of NOISY-GES and data splitting in terms of structural Hamming distance to true graph for varying n (x-axis) and d (y-axis). Left panel is the empty DAG setting; right panel is the sparse random DAG setting.

Validity Experiments. We quantify the severity of uncorrected inference after causal discovery by evaluating the probability of miscoverage of an effect of interest. In particular, we use the same data to both estimate the causal graph $\widehat{G}$ via GES and to compute a point estimate of the effect $\widehat{\beta}_{\widehat{G}}^{(i\to j)}$, and then we use a standard t-interval around the point estimate to produce a confidence region for the effect $\beta_{\widehat{G}}^{(i\to j)}$. We investigate two models for generating the true underlying graph: 1) an empty graph and 2) a sparse random graph. In either case, we first run GES to estimate a graph $\widehat{G}$. Then, we select an edge $e = X_i \to X_j$ uniformly over all edges in $\widehat{G}$ and compute a 95% confidence interval. We repeat this procedure 1000 times to estimate the probability of miscoverage of the population-level estimate $\beta_{\widehat{G}}^{(i\to j)}$ (which in the empty graph case is simply zero). In Figure 1 we plot the probability of error for varying sample size n and number of variables d.

Graph Recovery. In Figure 2 we compare GES and data splitting in terms of the structural Hamming distance (SHD) of their output to the true underlying graph. To implement a fair comparison between the two approaches, for a given max-information bound of NOISY-GES, we derive a splitting fraction p that makes the resulting confidence intervals of the same size (see Appendix). The blue entries correspond to NOISY-GES incurring lower SHD error and the red entries correspond to data splitting incurring lower SHD error, with the shade indicating the size of the difference. As we increase the number of variables, our algorithm consistently outperforms data splitting. Data splitting outperforms NOISY-GES in the lowest-dimensional setting ($d = 5$), which, as shown in Figure 1, coincides with the settings where inference after causal discovery is itself less problematic.

References

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *Annals of Statistics*, 41(2):802–837, 2013.

Andreas Buja, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544, 2019a.

Andreas Buja, Lawrence Brown, Arun Kumar Kuchibhotla, Richard Berk, Edward George, and Linda Zhao. Models as approximations II: A model-free theory of parametric regression. *Statistical Science*, 34(4):545–565, 2019b.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*, volume 9. Now Publishers, Inc., 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2350–2358, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 117–126, 2015b.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [No]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Supplementary Material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] The experiments were run locally on a personal computer.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Technical Lemmas

Lemma 1. *Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. random variables and $\{C_n\}_n$ a sequence of clipping thresholds such that $C_n \to \infty$. Then, $\frac{1}{n}\sum_{i=1}^n \min\{X_i, C_n\} \to_p \mathbb{E}X_1$.*

Proof. For any $\epsilon > 0$, by Chebyshev's inequality we have

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n \min\{X_i, C_n\} - \mathbb{E}\min\{X_1, C_n\}\right| \geq \epsilon\right\} \leq \frac{\mathrm{Var}(\min\{X_1, C_n\})}{n\epsilon^2} \leq \frac{\mathbb{E}X_1^2}{n\epsilon^2},$$

which tends to 0 as $n \to \infty$. Moreover, $\mathbb{E}\min\{X_1, C_n\} \to \mathbb{E}X_1$ by dominated convergence, hence we can conclude that $\frac{1}{n}\sum_{i=1}^n \min\{X_i, C_n\} \to_p \mathbb{E}X_1$. □

Lemma 2 (Closure under post-processing [Dwork et al., 2006])**.** *Let $\mathcal{A}(\cdot)$ be an ϵ-differentially private algorithm and let $\mathcal{B}$ be an arbitrary, possibly randomized map. Then, $\mathcal{B} \circ \mathcal{A}(\cdot)$ is ϵ-differentially private.*

Lemma 3 (Adaptive composition [Dwork et al., 2006])**.** *For $t \in [k]$, let $\mathcal{A}_t(\cdot, a_1, a_2, \ldots, a_{t-1})$ be ϵ_t-differentially private for all fixed $a_1, \ldots, a_{t-1}$. Then, the algorithm $\mathcal{A}_{\mathrm{comp}}$ which executes $\mathcal{A}_1, \ldots, \mathcal{A}_k$ in sequence and outputs $a_1 = \mathcal{A}_1(\mathcal{D}), a_2 = \mathcal{A}_2(\mathcal{D}, a_1), \ldots, a_k = \mathcal{A}_k(\mathcal{D}, a_1, \ldots, a_{k-1})$ is $(\sum_{t=1}^k \epsilon_t)$-differentially private.*

B Greedy Equivalence Search: Background

In this section we provide the details behind the greedy pass subroutine (Algorithm 3) that is used in GES. In particular, we review the definitions of valid $(\texttt{sgn})$-operators that appear in [Chickering, 2002], clarify what it means to apply a given operator to the current CPDAG, and explain how the score gains $\Delta S^{\texttt{sgn}}(e, \widehat{G}, \mathcal{D})$ are computed. As before, we use $\widehat{G}$ to denote the CPDAG maintained by GES.

Before we define $(\texttt{sgn})$-operators, we briefly review some graph-theoretic preliminaries. We say two nodes X_a, X_b are *neighbors* in a CPDAG $\widehat{G}$ if they are connected by an undirected edge, and *adjacent* if they are connected by any edge (directed or undirected). We also call a path from X_a to X_b in a CPDAG *semi-directed* if each edge along it is either undirected or directed away from X_a.

Definition 6. *For non-adjacent X_a and X_b in $\widehat{G}$, and a subset $\mathbf{T}$ of X_b's neighbors that are not adjacent to X_a, the* $\mathrm{Insert}(X_a, X_b, \mathbf{T})$ *operator is defined as the procedure that modifies $\widehat{G}$ by inserting edge $X_a \to X_b$ and for each $T \in \mathbf{T}$, converting $T - X_b$ to $T \to X_b$.*

Definition 7. *For X_a and X_b in $\widehat{G}$ connected as $X_a - X_b$ or $X_a \to X_b$, and a subset $\mathbf{T}$ of X_b's neighbors that are adjacent to X, the* $\mathrm{Delete}(X_a, X_b, \mathbf{T})$ *operator is defined as the procedure that modifies $\widehat{G}$ by deleting the edge between X_a and X_b, and for each $T \in \mathbf{T}$, converting $X_b - T$ to $X_b \to T$ and $X_a - T$ to $X_a \to T$.*

We use "$(+)$-operator" (resp. "$(-)$-operator") as a shorthand for the Insert operator (resp. the Delete operator).

Now that we have a definition of $(\texttt{sgn})$-operators, we need to define which operators are *valid* to apply to the current graph. For example, if we were greedily updating only a single DAG and not a CPDAG, we would only consider edge additions that maintain the DAG structure. We define an analogous form of validity for CPDAGs, which requires a bit more care. Let $\mathbf{NA}_{X_b, X_a}$ be the neighbors of X_b that are adjacent to X_a.

Definition 8. *We say that* $\mathrm{Insert}(X_a, X_b, \mathbf{T})$ *is valid if $\mathbf{NA}_{X_b, X_a} \cup \mathbf{T}$ is a clique and every semi-directed path from X_b to X_a contains a node in $\mathbf{NA}_{X_b, X_a} \cup T$.*

Definition 9. *We say that* $\mathrm{Delete}(X_a, X_b, \mathbf{T})$ *is valid if $\mathbf{NA}_{X_b, X_a} \setminus \mathbf{T}$ is a clique.*

For a valid $(\texttt{sgn})$-operator, Chickering also defines how to properly score the gain due to applying it. In particular, the score gain due to executing $\mathrm{Insert}(X_a, X_b, \mathbf{T})$ is defined as $\Delta S^{+}((X_a, X_b, \mathbf{T}), \widehat{G}, \mathcal{D}) = s(X_a, \mathbf{NA}_{X_b, X_a} \cup \mathbf{T} \cup \mathbf{Pa}_{X_b} \cup X_a, \mathcal{D}) - s(X_b, \mathbf{NA}_{X_b, X_a} \cup \mathbf{T} \cup \mathbf{Pa}_{X_b} \cup X_a, \mathcal{D})$.

This expression is essentially an application of the formula decomposition of the score gain for a specific DAG G consistent with the CPDAG $\widehat{G}$ and edge $e = (X_a \to X_b)$. Similarly, the score gain due to executing $\mathrm{Delete}(X_a, X_b, \mathbf{T})$ is defined as $\Delta S^{-}((X_a, X_b, \mathbf{T}), \widehat{G}, \mathcal{D}) = s(X_b, \{\mathbf{NA}_{X_b, X_a} \setminus \mathbf{T}\} \cup \{\mathbf{Pa}_{X_b} \setminus X_a\}, \mathcal{D}) - s(X_b, \{\mathbf{NA}_{X_b, X_a} \setminus \mathbf{T}\} \cup \mathbf{Pa}_{X_b}, \mathcal{D})$.

Having laid out this preamble, we can now state more precisely the greedy pass subroutine (Algorithm 3) of noisy GES.

Algorithm 4 GreedyPass

input : initial graph $\widehat{G}_0$, data set $\mathcal{D}$, maximum number of edges $E_{\max}$, score S with local score sensitivity τ, privacy parameters $\epsilon_{\text{score}}, \epsilon_{\text{thresh}}$, pass indicator $\texttt{sgn} \in \{+, -\}$

output: estimated causal graph $\widehat{G}$

Initialize $\widehat{G} \leftarrow \widehat{G}_0$

Sample noisy threshold $\nu \sim \text{Lap}\left(\frac{4\tau}{\epsilon_{\text{thresh}}}\right)$

for $t = 1, 2, \ldots, E_{\max}$ **do**

  if $sgn = +$ **then**

    Construct set $\mathcal{E}_t^{+}$ of all valid $\mathrm{Insert}(X_a, X_b, \mathbf{T})$ operators (Def. 8)

  else if $sgn = -$ **then**

    Construct set $\mathcal{E}_t^{-}$ of all valid $\mathrm{Delete}(X_a, X_b, \mathbf{T})$ operators (Def. 9)

  end

  For all $e \in \mathcal{E}_t^{\texttt{sgn}}$, compute $\Delta S^{\texttt{sgn}}(e, \widehat{G}, \mathcal{D})$ and sample $\xi_{t,e} \overset{\text{i.i.d.}}{\sim} \text{Lap}\left(\frac{4\tau}{\epsilon_{\text{score}}}\right)$

  Set $e_t^* = \arg\max_{e \in \mathcal{E}_t^{\texttt{sgn}}} \Delta S^{\texttt{sgn}}(e, \widehat{G}, \mathcal{D}) + \xi_{t,e}$

  Sample $\eta_t \sim \text{Lap}\left(\frac{8\tau}{\epsilon_{\text{thresh}}}\right)$

  if $\Delta S^{sgn}(e_t^*, \widehat{G}, \mathcal{D}) + \eta_t > \nu$ **then**

    Apply operator e_t^* to $\widehat{G}$

  else

    break

  end

end

Return $\widehat{G}$

C Noisy Causal Graph Discovery: Proofs

C.1 Proof of Proposition 2

The proposition is an application of the privacy guarantees of the Report Noisy Max mechanism in differential privacy (see, for example, Chapter 3.3 in the book [Dwork and Roth, 2014]). In addition, the privacy analysis of Algorithm 2 strictly subsumes the privacy analysis of Algorithm 1.

C.2 Proof of Theorem 1

By Proposition 1, we can bound the max-information between $\widehat{G}$ and $\mathcal{D}$:

$$\mathcal{I}_\infty^\beta(\widehat{G}; \mathcal{D}) \leq \frac{n}{2}\epsilon^2 + \epsilon\sqrt{n\log(2/\beta)/2}.$$

The definition of max-information, in turn, implies that

$$\begin{aligned}
&\mathbb{P}\Big\{\exists(i,j) \in \mathcal{I}_G : \beta_G^{(i\to j)} \notin \mathrm{CI}_G^{(i\to j)}(\tilde{\alpha}), \widehat{G} = G\Big\} \\
&\quad \leq \exp\left(\mathcal{I}_\infty^\beta(\widehat{G}; \mathcal{D})\right) \mathbb{P}\Big\{\exists(i,j) \in \mathcal{I}_G : \beta_G^{(i\to j)} \notin \mathrm{CI}_G^{(i\to j)}(\tilde{\alpha}; \tilde{\mathcal{D}}), \widehat{G} = G\Big\} \\
&\quad \leq \exp\left(\frac{n}{2}\epsilon^2 + \epsilon\sqrt{n\log(2/\beta)/2}\right)\tilde{\alpha} \\
&\quad = \alpha.
\end{aligned}$$

Marginalizing over all graphs G yields the final theorem statement.

D Noisy Greedy Equivalence Search: Proofs

D.1 Proof of Proposition 3

As mentioned earlier, the proof relies on the analysis of two differentially private mechanisms: Report Noisy Max and AboveThreshold [Dwork and Roth, 2014]. To facilitate the proof, in Algorithm 5 we provide an equivalent reformulation of Algorithm 3 that allows decoupling the analyses of these two mechanisms.

Algorithm 5 Decoupled GreedyPass

input : initial graph $\widehat{G}_0$, data set $\mathcal{D}$, maximum number of edges $E_{\max}$, score S with local score sensitivity τ, privacy parameters $\epsilon_{\text{score}}, \epsilon_{\text{thresh}}$, pass indicator $\texttt{sgn} \in \{+, -\}$

output: estimated causal graph $\widehat{G}$

Initialize $\widehat{G} \leftarrow \widehat{G}_0$

Get potential operators $\mathcal{E} \leftarrow$ ProposeOperators$(\widehat{G}, \mathcal{D}, E_{\max}, S, \tau, \epsilon_{\text{score}}, \texttt{sgn})$

Get selected operator subset $\mathcal{E}^* \leftarrow$ SelectOperators$(\widehat{G}, \mathcal{D}, S, \tau, \epsilon_{\text{thresh}}, \texttt{sgn}, \mathcal{E})$

for $t = 1, \ldots, |\mathcal{E}^*|$ **do**

  Apply e_t^* to $\widehat{G}$

end

Return $\widehat{G}$

Algorithm 6 ProposeOperators

input : initial graph $\widehat{G}_0$, data set $\mathcal{D}$, maximum number of edges $E_{\max}$, score S with local score sensitivity τ, privacy parameter ϵ_{score}, pass indicator $\texttt{sgn} \in \{+, -\}$
output: proposed set of operators $\mathcal{E}$

Initialize $\widehat{G} \leftarrow \widehat{G}_0$
Initialize $\mathcal{E} \leftarrow \varnothing$
for $t = 1, 2, \ldots, E_{\max}$ **do**
  Construct set $\mathcal{E}_t^{\text{sgn}}$ of valid $(\texttt{sgn})$-operators
  For all $e \in \mathcal{E}_t^{\text{sgn}}$, compute $\Delta S^{\text{sgn}}(e, \widehat{G}, \mathcal{D})$ and sample $\xi_{t,e} \overset{\text{i.i.d.}}{\sim} \text{Lap}\left(\frac{4\tau}{\epsilon_{\text{score}}}\right)$
  Set $e_t = \arg\max_{e \in \mathcal{E}_t^{\text{sgn}}} \Delta S^{\text{sgn}}(e, \widehat{G}, \mathcal{D}) + \xi_{t,e}$
  Add operator e_t to $\mathcal{E}$
  Apply operator e_t to $\widehat{G}$
end
Return $\mathcal{E} = (e_1, \ldots, e_{E_{\max}})$

Algorithm 7 SelectOperators

input : initial graph $\widehat{G}_0$, data set $\mathcal{D}$, score S with local score sensitivity τ, privacy parameter ϵ_{thresh}, pass indicator $\texttt{sgn} \in \{+, -\}$, set of proposed operators $\mathcal{E}$
output: set of operators $\mathcal{E}^*$

Sample noisy threshold $\nu \sim \text{Lap}\left(\frac{4\tau}{\epsilon_{\text{thresh}}}\right)$
Initialize $\mathcal{E}^* \leftarrow \varnothing$
Initialize $\widehat{G} \leftarrow \widehat{G}_0$
for $t = 1, 2, \ldots, |\mathcal{E}|$ **do**
  Sample $\eta_t \sim \text{Lap}\left(\frac{8\tau}{\epsilon_{\text{thresh}}}\right)$
  if $\Delta S^{\text{sgn}}(e_t, \widehat{G}, \mathcal{D}) + \eta_t \geq \nu$ **then**
    Add e_t^* to $\mathcal{E}^*$
    Apply e_t^* to $\widehat{G}$
  else
    break
  end
end
Return $\mathcal{E}^* = (e_1^*, e_2^*, \ldots)$

We argue that the two subroutines composed in the greedy pass, namely ProposeOperators (Algorithm 6) and SelectOperators (Algorithm 7), are differentially private. By the closure of differential privacy under post-processing (Lemma 2), this will imply that Algorithm 5, which returns $\widehat{G}$, is also differentially private, since $\widehat{G}$ is merely a post-processing of the selected operators $\mathcal{E}^*$.

The privacy guarantee of Algorithm 6 is implied by the usual privacy guarantee of Report Noisy Max and composition of differential privacy. Note that the construction of the set $\mathcal{E}_t^{\text{sgn}}$ at every time step is only a function of the current graph $\widehat{G}$ and not of the data, i.e. it is independent of the data conditioned on $\widehat{G}$. Formally, the key component is the following lemma:

Lemma 4. *For any $t \in [E_{\max}]$, selecting e_t is ϵ_{score}-differentially private; that is, for any operator $e_0 \in \mathcal{E}_t^{sgn}$, it holds that*

$$\mathbb{P}\left\{ \arg\max_{e \in \mathcal{E}_t^{sgn}} \Delta S^{sgn}(e, \widehat{G}, \mathcal{D}) + \xi_{t,e} = e_0 \;\middle|\; \widehat{G} \right\} \leq e^{\epsilon_{\text{score}}} \mathbb{P}\left\{ \arg\max_{e \in \mathcal{E}_t^{sgn}} \Delta S^{sgn}(e, \widehat{G}, \mathcal{D}') + \xi_{t,e} = e_0 \;\middle|\; \widehat{G} \right\}.$$

for any current graph $\widehat{G}$ and any two neighboring data sets $\mathcal{D}, \mathcal{D}'$.

Proof. Denote $r_e \doteq \Delta S^{\text{sgn}}(e, \widehat{G}, \mathcal{D})$ and $r'_e \doteq \Delta S^{\text{sgn}}(e, \widehat{G}, \mathcal{D}')$. For a fixed $e_0 \in \mathcal{E}_t^{\text{sgn}}$, define

$$\xi^\star_{t,e_0} \doteq \min\{\xi : r_{e_0} + \xi > r_{e'} + \xi_{t,e'} \; \forall e' \neq e_0\}.$$

For fixed $\{\xi_{t,e'}\}_{e' \neq e_0}$, we have that e_0 will be the selected operator on $\mathcal{D}$ if and only if $\xi_{t,e_0} \geq \xi^\star_{t,e_0}$. Further, by the bounded sensitivity of the local scores, we have that for all $e' \neq e_0$:

$$\begin{aligned} & r_{e_0} + \xi^\star_{t,e_0} > r_{e'} + \xi_{t,e'} \\ \Rightarrow\ & r'_{e_0} + 2\tau + \xi^\star_{t,e_0} > r'_{e'} - 2\tau + \xi_{t,e'} \\ \Rightarrow\ & r'_{e_0} + \left(4\tau + \xi^\star_{t,e_0}\right) > r_{e'} + \xi_{t,e'}. \end{aligned}$$

Therefore, as long as $\xi_{t,e_0} \geq 4\tau + \xi^\star_{t,e_0}$, the selection on $\mathcal{D}$ will be e_0 as well. Using the form of the density of $\xi_{t,e_0} \sim \text{Lap}\left(\frac{4\tau}{\epsilon_{\text{score}}}\right)$, we have that:

$$\begin{aligned} \mathbb{P}\left\{ \arg\max_e r'_e + \xi_{t,e} = e_0 \mid \{\xi_{t,e'}\}_{e' \neq e_0} \;\middle|\; \widehat{G} \right\} &\geq \mathbb{P}\left\{ \xi_{t,e_0} \geq 4\tau + \xi^\star_{t,e_0} \right\} \\ &\geq e^{-\epsilon_{\text{score}}} \mathbb{P}\left\{ \xi_{t,e_0} \geq \xi^\star_{t,e_0} \right\} \\ &= \mathbb{P}\left\{ \arg\max_e r_e + \xi_{t,e} = e_0 \mid \{\xi_{t,e'}\}_{e' \neq e_0} \;\middle|\; \widehat{G} \right\} \end{aligned}$$

By taking iterated expectations, overall we have

$$\mathbb{P}\left\{ \arg\max_{e \in \mathcal{E}_t^{\text{sgn}}} \Delta S^{\text{sgn}}(e, \widehat{G}, \mathcal{D}) + \xi_{t,e} = e_0 \;\middle|\; \widehat{G} \right\} \leq e^{\epsilon_{\text{score}}} \mathbb{P}\left\{ \arg\max_{e \in \mathcal{E}_t^{\text{sgn}}} \Delta S^{\text{sgn}}(e, \widehat{G}, \mathcal{D}') + \xi_{t,e} = e_0 \;\middle|\; \widehat{G} \right\}$$

for all neighboring data sets $\mathcal{D}, \mathcal{D}'$, ensuring the desired privacy. □

This directly implies the following result:

Lemma 5 (Privacy of ProposeOperators)**.** *Algorithm 6 is $E_{\max}\epsilon_{\text{score}}$-differentially private.*

Proof. The result follows directly from Lemma 4, by applying the adaptive composition rule for differential privacy (Lemma 3) over $E_{\max}$ steps. □

Now we isolate the second component of the greedy pass: checking if the operator's contribution is positive. To analyze this component independently of the selection of potential operators, we consider Algorithm 7 which receives a set of proposed operators $\mathcal{E}$ and outputs only the first $E^*_{\max} \leq E_{\max}$ of them which pass the noisy threshold test. Note that $E^*_{\max}$ is random and data-dependent.

In what follows we use $\mathcal{E}^*(\mathcal{D})$ and $\mathcal{E}^*(\mathcal{D}')$ to denote the output of Algorithm 7 on two neighboring data sets $\mathcal{D}, \mathcal{D}'$.

Lemma 6 (Privacy of SelectOperators)**.** *For any input set of proposed edges $\mathcal{E} = (e_1, \ldots, e_{E_{\max}})$, Algorithm 7 is ϵ_{thresh}-differentially private; that is, for any $1 \leq k \leq E_{\max} + 1$:*

$$\mathbb{P}\{\mathcal{E}^*(\mathcal{D}) = (e_j)_{j<k}\} \leq e^{\epsilon_{\text{thresh}}} \mathbb{P}\{\mathcal{E}^*(\mathcal{D}') = (e_j)_{j<k}\}$$

given any two neighboring data sets $\mathcal{D}, \mathcal{D}'$.

Proof. Fix $1 \leq k \leq E_{\max} + 1$ and consider $(e_1, \ldots, e_k)$. Let $G_1, \ldots, G_k$ be the graphs resulting from the application of operators e_t in sequence, starting from the initial graph $\widehat{G}_0$. Define $r_t = \Delta S^{\text{sgn}}(e_t, G_{t-1}, \mathcal{D})$ and $r'_t = \Delta S^{\text{sgn}}(e_t, G_{t-1}, \mathcal{D}')$. Condition on $\eta_1, \ldots, \eta_{k-1}$ and define the following quantity that captures the minimal value of the noisy score gain up to time $k-1$:

$$g(\mathcal{D}) = \min_{i<k}\{r_i + \eta_i\},$$

and analogously for $\mathcal{D}'$:

$$g(\mathcal{D}') = \min_{i<k}\{r'_i + \eta_i\}.$$

Using these quantities we can directly express the probability of outputting exactly the first $k-1$ proposed operators, i.e. breaking at the k-th step of the algorithm:

$$\begin{aligned} \mathbb{P}\{\mathcal{E}^*(\mathcal{D}) = (e_j)_{j<k}\} &= \mathbb{P}\{\nu \in (r_k + \eta_k, g(\mathcal{D}))]\} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p_{\eta_k}(q) p_\nu(w) \mathbf{1}\{w \in (r_k + q, g(\mathcal{D}))]\} dq dw. \end{aligned}$$

With the change of variables $q' = q - g(\mathcal{D}) + g(\mathcal{D}') + r_k - r'_k$, $w' = w + g(\mathcal{D}') - g(\mathcal{D})$, we have

$$\mathbf{1}\{w \in (r_k+q, g(\mathcal{D})]\} = \mathbf{1}\{w'+g(\mathcal{D})-g(\mathcal{D}') \in (q'+g(\mathcal{D})-g(\mathcal{D}')+r'_k, g(\mathcal{D})]\} = \mathbf{1}\{w' \in (r'_k+q', g(\mathcal{D}')]\}$$

and thus

$$\begin{aligned}&\mathbb{P}\{\mathcal{E}^*(\mathcal{D}) = (e_j)_{j<k}\}\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p_{\eta_k}(q' + g(\mathcal{D}) - g(\mathcal{D}') - r_k + r'_k)p_\nu(w' - g(D') + g(\mathcal{D}))\mathbf{1}\{w' \in (r'_k + q', g(\mathcal{D}')]\}dq'dw'.\end{aligned}$$

Observe that r_t is 2τ-sensitive since the local scores are τ-sensitive, and hence $g(\mathcal{D})$ is 2τ-sensitive as well. This implies that $|q' - q| \leq 4\tau$, $|w' - w| \leq 2\tau$, so by the form of the Laplace density we have

$$p_{\eta_k}(q' + g(\mathcal{D}) - g(\mathcal{D}') - r_k + r'_k) \leq e^{\epsilon_{\text{thresh}}/2}p_{\eta_k}(q'),\ \ p_\nu(w' - g(D') + g(\mathcal{D})) \leq e^{\epsilon_{\text{thresh}}/2}p_\nu(w').$$

Putting everything together, we have:

$$\begin{aligned}\mathbb{P}\{\mathcal{E}^*(\mathcal{D}) = (e_j)_{j<k}\} &\leq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{\epsilon_{\text{thresh}}/2}p_{\eta_k}(q')p_\nu(w')e^{\epsilon_{\text{thresh}}/2}\mathbf{1}\{w' \in (r'_k + q', g(\mathcal{D}')]\}dq'dw'\\ &= e^{\epsilon_{\text{thresh}}}\mathbb{P}\{\mathcal{E}^*(\mathcal{D}') = (e_j)_{j<k}\},\end{aligned}$$

which is the desired guarantee. □

Finally, we combine the guarantees of Lemma 4 and Lemma 6 to infer the privacy parameter of Decoupled GreedyPass (Algorithm 5), which is equivalent to GreedyPass from Algorithm 2. The following statement follows from a direct application of privacy composition (i.e., Lemma 3).

Lemma 7 (Privacy of Decoupled GreedyPass)**.** *Algorithm 5 is $\epsilon_{\text{thresh}} + E_{\max}\epsilon_{\text{score}}$-differentially private.*

Proof of Proposition 3. Since the GES algorithm (Algorithm 2) consists of two executions of Greedy-Pass, which is equivalent to the Decoupled GreedyPass, we can apply Lemma 7 and Lemma 3 to conclude that GES is $(2\epsilon_{\text{thresh}} + 2E_{\max}\epsilon_{\text{score}})$-differentially private. □

D.2 Proof of Proposition 4

We show that, in the large sample limit, private GES behaves identically to the standard GES method. Denote by $e_1^*, e_2^*, \ldots$ the insertion operators selected by non-private GES in the forward greedy pass and by $\widehat{G}_t$ the CPDAG constructed at the end of step t of the forward pass. Further, let $\texttt{Gap} = \min_t \min_{e \neq e_t^*} \Delta S^+(e_t^*, \widehat{G}_{t-1}, \mathcal{D}) - \Delta S^+(e, \widehat{G}_{t-1}, \mathcal{D})$. In words, Gap is the gap in score improvement between the optimal edge at time t and the second best edge at time t, minimized over all steps t. Notice that by the existence of distinct $\Delta s^{\texttt{sgn}}_{e,G}$, we know that $\lim_{n\to\infty} \texttt{Gap} > 0$. Moreover, $\frac{\tau}{\epsilon_{\text{score}}}, \frac{\tau}{\epsilon_{\text{thresh}}} = o(1)$ implies that the noise level vanishes asymptotically. Putting all of this together implies that the limiting probability that noisy GES selects $e_1^*, e_2^*, \ldots$ is one. By a similar argument we conclude that noisy GES halts the forward phase at the same step as the non-noisy GES, and thus we have shown that the output of the forward pass of noisy GES is asymptotically the same as that of non-noisy GES. By an analogous argument it follows that the outputs of the backward pass are identical, which completes the proof of the first claim. The second claim follows directly by putting together the first claim and the classical consistency result for GES [Chickering, 2002].

D.3 Proof of Proposition 5

First we prove that the score is $\frac{C}{n\sigma^2}$-sensitive. The proof generalizes the proof of Claim **??**. Let $\mathcal{D} = \{X^{(k)}\}_{k=1}^n$ and $\mathcal{D}' = \{X'^{(k)}\}_{k=1}^n$ denote two data sets that differ in one entry, and without loss

of generality assume they differ in the first entry. Let j index an arbitrary variable and denote

$$\theta_{\mathcal{D}} = \arg\min_{\theta} L_j(\theta, \mathcal{D}) := \arg\min_{\theta} \frac{1}{n\sigma^2} \sum_{k=1}^{n} \min\left\{ \left(X_j^{(k)} - \sum_{s \in \mathbf{Pa}_j^G} \theta_s X_s^{(k)} \right)^2, C \right\},$$

$$\theta_{\mathcal{D}'} = \arg\min_{\theta} L_j(\theta, \mathcal{D}') := \arg\min_{\theta} \frac{1}{n\sigma^2} \sum_{k=1}^{n} \min\left\{ \left(X_j^{'(k)} - \sum_{s \in \mathbf{Pa}_j^G} \theta_s X_s^{'(k)} \right)^2, C \right\}.$$

We argue that $|L_j(\theta_{\mathcal{D}}, \mathcal{D}) - L_j(\theta_{\mathcal{D}'}, \mathcal{D}')| \leq \frac{C}{n\sigma^2}$. First, for all θ, $|L_j(\theta, \mathcal{D}) - L_j(\theta, \mathcal{D}')| \leq \frac{C}{n\sigma^2}$ since the corresponding sums only differ in one entry. Combining this fact with the optimality condition for $\theta_{\mathcal{D}}$, we get

$$L_j(\theta_{\mathcal{D}}, \mathcal{D}) \leq L_j(\theta_{\mathcal{D}'}, \mathcal{D}) \leq L_j(\theta_{\mathcal{D}'}, \mathcal{D}') + \frac{C}{n\sigma^2}.$$

Analogously we obtain that $L_j(\theta_{\mathcal{D}'}, \mathcal{D}') \leq L_j(\theta_{\mathcal{D}}, \mathcal{D}) + \frac{C}{n\sigma^2}$, which completes the proof of the first claim.

The proof of local consistency directly relies on local consistency of the standard BIC score, in combination with Lemma 1. In particular, Lemma 1 implies that

$$\frac{1}{n\sigma^2} \sum_{k=1}^{n} \min\left\{ \left(X_j^{(k)} - \sum_{s \in \mathbf{Pa}_j^G} \theta_s X_s^{(k)} \right)^2, C \right\} \to_p \frac{1}{\sigma^2} \mathbb{E}\left(X_j^{(1)} - \sum_{s \in \mathbf{Pa}_j^G} \theta_s X_s^{(1)} \right)^2,$$

since $C = \omega(1)$. In other words, the asymptotic behavior of the clipped BIC score is identical to the usual BIC score, whenever the clipping threshold diverges.

Therefore, for any G and candidate edge $X_i \to X_j$ such that $X_j \not\perp X_i | X_{\mathbf{Pa}_j^G}$ we have

$$\begin{aligned}
\lim_{n\to\infty} s_{\text{BIC}}^C(X_j, \mathbf{Pa}_j^G \cup X_i, \mathcal{D}) &= \lim_{n\to\infty} s_{\text{BIC}}(X_j, \mathbf{Pa}_j^G \cup X_i, \mathcal{D}) \\
&> \lim_{n\to\infty} s_{\text{BIC}}(X_j, \mathbf{Pa}_j^G, \mathcal{D}) \\
&= \lim_{n\to\infty} s_{\text{BIC}}^C(X_j, \mathbf{Pa}_j^G, \mathcal{D}),
\end{aligned}$$

where the second step follows by local consistency of the standard BIC score, and the first and final steps follow by the condition that $C = \omega(1)$. This proves the first condition of Definition **??**.

The proof of the second condition follows analogously; suppose $X_j \perp X_i | X_{\mathbf{Pa}_j^G}$, then

$$\begin{aligned}
\lim_{n\to\infty} s_{\text{BIC}}^C(X_j, \mathbf{Pa}_j^G \cup X_i, \mathcal{D}) &= \lim_{n\to\infty} s_{\text{BIC}}(X_j, \mathbf{Pa}_j^G \cup X_i, \mathcal{D}) \\
&< \lim_{n\to\infty} s_{\text{BIC}}(X_j, \mathbf{Pa}_j^G, \mathcal{D}) \\
&= \lim_{n\to\infty} s_{\text{BIC}}^C(X_j, \mathbf{Pa}_j^G, \mathcal{D}),
\end{aligned}$$

which completes the proof.