# Markov-Boundary Consistent Feature Attribution

**Mateusz Gajewski** [1 2] **Mateusz Olko** [3 2] **Mikołaj Morzy** [1] **Piotr Sankowski** [3 4 5]

## Abstract

Feature attribution methods aim to explain the predictions of machine learning models by assigning importance scores to input features. Recent work has highlighted the importance of developing attribution methods that respect causal structures. Furthermore, they showed that existing approaches can assign significant importance to variables outside the Markov boundary, even though these variables provide no additional predictive information when the Markov boundary is observed. To address these limitations we design a new attribution method that accounts for both predictive power and causal structure of the features. Our method does not assume access to the structure and achieves balanced attributions using properly defined characteristic function. We show that our method provably assigns high attributions to the variables in the Markov boundary and experimentally evaluate it in a fairness inspired setting.

## 1. Introduction

Feature attribution methods aim to explain the predictions of machine learning models by assigning importance scores to input features. In high-stakes domains such as healthcare and finance, such explanations are crucial for ensuring transparency and accountability.

The intersection of feature attribution and causal inference has emerged as a critical area of research in explainable AI. Recent work has highlighted the importance of developing attribution methods that respect causal structures, moving beyond purely correlational approaches to explainability. Several approaches have been proposed to align feature attributions with causal principles. Frye et al. (2020) incorporate causal relationships by modifying the standard Shapley framework to respect the directionality of effects. Similarly, Causal Shapley Values (Heskes et al., 2020) adapt the value function to reflect causal rather than purely statistical contributions. The work of Janzing et al. (2024) offers another insightful perspective on how variables contribute to predictions through causal mechanisms.

Our work is motivated by a key limitation in existing attribution approaches. Ma & Tourani (2020) demonstrated that standard Shapley values are inconsistent with respect to the Markov boundary of the target variable. Specifically, they showed that Shapley values can assign significant importance to variables outside the Markov boundary, even though these variables provide no additional predictive information when the Markov boundary is observed.

We consider problem of attributing nodes in a structural causal model (SCM). From causal theory, variables in the Markov boundary are sufficient for optimal prediction of the target variable. However, variables outside the Markov boundary still exhibit predictive power when considered individually.

This creates a tension: while Markov boundary variables are theoretically sufficient when observed together, non-Markov boundary variables retain predictive utility. Thus we argue that complete exclusion of these variables from attribution would ignore their actual predictive contribution. A principled attribution method should therefore satisfy two properties: non-Markov boundary variables can receive positive attribution when they contribute predicatively, but the sum of attributions to Markov boundary variables should exceed the sum of attributions to variables outside the boundary.

Our approach bridges the gap between causal inference and feature attribution, providing explanations that are both predicatively relevant and causally meaningful. Specifically, our contributions include:

- We design a new attribution method based on explained variance that has desirable properties with respect to the Markov boundary of the explained variable.
- We show that our method provably attributes higher scores to Markov boundary of the explained variable

[*]Equal contribution [1] Faculty of Computing and Telecommunications Poznan University of Technology, Poznan, Poland [2]IDEAS NCBR, Warsaw, Poland [3]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland [4]Research Institute IDEAS, Warsaw, Poland [5]MIM Solutuions, Warsaw, Poland. Correspondence to: Mateusz Gajewski <mg96272@gmail.com>.

*Proceedings of the $42^{nd}$ International Conference on Machine Learning*, Vancouver, Canada. PMLR 267, 2025. 

in linear models with additive gaussian noise.

- We demonstrated desirable properties of our method on a synthetic dataset motivated by fairness analysis use case.

## 2. Preliminaries

### 2.1. Causal Models

**Additive noise models (ANM).** Consider a directed acyclic graph (DAG) $G = (V, E)$ representing the causal structure among variables. A structural additive noise model over graph $G$ is defined by a set of equations:

$$X_i = f(Pa_G(X_i)) + \epsilon_i, \tag{1}$$

where $Pa_G(X_i)$ denotes the set of parents of $X_i$ in the graph $G$, $\epsilon_i \sim \mathcal{N}_i$ is a noise term of node $i$, and $\mathcal{N}_i$ are independent noise distributions. When $f$ is linear and $N_i$ are gaussian we call such setup a linear model with additive gaussian noise.

**Markov Boundary** When predicting random variable based on other variables in the graph usually only a subset of the variables is required. We can formalize this observation. When predicting variable $Y$ in a random variable set $X = \{X_i, ..., X_n\}$, there exists a set $S \subseteq X$ that renders all others variables independent from $Y$: $Y \perp\!\!\!\perp X \setminus S \mid S$. The minimal of such sets is called *the Markov boundary* (Pearl, 2009).

In case of additive noise models the Markov boundary consists of: parents of $Y$, children of $Y$, and other parents of the children of $Y$. Please see Figure 1.

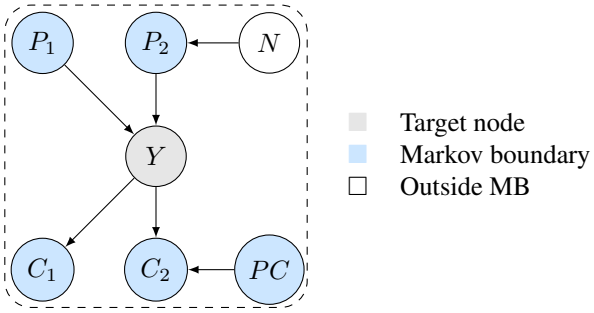


Figure 1: Example of a Markov boundary for node $Y$, including parents ($P_1$, $P_2$), children ($C_1$, $C_2$), and parents of children ($PC$). Node $N$ is outside the Markov boundary.

### 2.2. Cooperative Game Theory

**Characteristic Function** The feature attribution problem is often formulated within the framework of cooperative game theory. (Lundberg & Lee, 2017) (Štrumbelj & Kononenko, 2011) In this setting attributions are obtained as a solution to the game defined by:

- A set of players $N = \{1, 2, \ldots, n\}$
- A characteristic function $v : 2^N \to \mathbb{R}$ that assigns a real value to each coalition $S \subseteq N$, with $v(\emptyset) = 0$

In the feature attribution context, players $N$ correspond to the random variables $X$, and the characteristic function measures the predictive power of subsets of $X$. Intuitively, $v(S)$ represents the sum of attributions that subset $S \in X$ can achieve. The precise definition of the characteristic function is one of our contributions and will be described in the next section.

**Core of the Game** The core represents a set of feature attributions where the total payoff is exactly distributed among all variables in $X$ and no subset $S \in X$ can break away and achieve better payoffs on their own. Intuitively the core represents stable attribution allocations where no subset of variables has an incentive to form their own coalition, since each group receives at least as much attribution as they could achieve independently.

However, the core may be empty for some games. The existence of the core depends on properties of the characteristic function. One important property that ensures a non-empty core is game convexity. We say that a cost cooperative game $(N, v)$ is convex if for all $i \in N$ and all $S \subseteq T \subseteq N \setminus \{i\}$ (Shapley, 1971):

$$v(S \cup \{i\}) - v(S) \geq v(T \cup \{i\}) - v(T) \tag{2}$$

## 3. A New Attribution Method

In this section we describe our feature attribution method. We begin by describing the characteristic function of the cooperative game that provides the feature attributions. Then we follow with theoretical results for linear models with additive gaussian noise where we show that sum of attributions to variables in the Markov boundary of the explained variables exceeds the sum of attributions to other variables. We provide high-level proof sketches for theorems in the main text. The full proofs can be found in the Appendix A.

### 3.1. Characteristic Function Based on Explained Variance

We aim to explain a predictive model $f$ that, we assume was optimized to approximate $\mathbb{E}[Y|X]$. For example for linear gausian data such a predictive model is equivalent to the model trained using MSE (Hastie et al., 2009).

To formulate feature attribution as a cooperative game, we need to define a characteristic function that measures the

contribution of each feature coalition. To do this we use two components: (i) the explained variance and (ii) the penalty that counts the number of supersets of S that improve the explained variance, see Equation 5.

**Variance of Expected Value** For a given instance $x$ and a subset of features $S$, we describe the first part of proposed characteristic function as a variance of expected value of $Y$ given $x_S$:

$$\text{Var}(\mathbb{E}[Y|x_S]) = E_{X_{-S}|x_S}[Var(E[Y|x_S, X_{-S}])] = E_{X_{-S}|x_S}[f(x)] \quad (3)$$

The choice of variance as our characteristic function is motivated by several desirable properties like Translation Invariance (is invariant to constant shifts) and symmetry (treats positive and negative deviations equally, avoiding bias toward either direction).

**The penalty function** We introduce a penalty function $w(S)$ that counts the number of supersets of $S$ that decrease the variance:

$$w(S) = |\{S' \subseteq V \setminus \{Y\} : S \subseteq S' \text{ and } \text{Var}(\mathbb{E}[Y|x_{S'}]) < \text{Var}(\mathbb{E}[Y|x_S])\}| \quad (4)$$

The penalty function ensures desirable properties of attribution with respect to the Markov boundary of $Y$. The full characteristic function for a coalition $S$ is defined as:

$$v(S) = -\text{Var}(\mathbb{E}[Y|x_S]) + a \cdot w(S) \quad (5)$$

where $a$ is a penalization coefficient.

### 3.2. Theoretical Results

We now present the main theoretical results for our new attribution method. First, we provide existence results for our the valid attributions. Than, we describe the relationship between the attributions and causal structure.

**Theorem 3.1** (Existence of Core Solutions)**.** *Consider the cooperative game with characteristic function $v(S) = -Var(\mathbb{E}[Y|x_S]) + a \cdot w(S)$ defined on $X$. There exists a threshold $a^* \leq 0$ such that for any $a < a^*$, the game is convex and therefore has a non-empty core.*

This result guarantees that by appropriately choosing the coefficient $a$, we can ensure that stable attributions exist for our feature attribution problem. The convexity property ensures that the core is non-empty and that allocation methods like the Shapley values lie within the core.

*Proof sketch of Theorem 3.1.* For a cooperative game to have a non-empty core, convexity of the characteristic function is sufficient. We show that for sufficiently negative values of $a$, our modified characteristic function becomes convex.

Let $S \subseteq T \subseteq V \setminus \{Y\}$ and $i \in V \setminus \{Y \cup T\}$. For convexity, we need:

$$v(S \cup \{i\}) - v(S) \geq v(T \cup \{i\}) - v(T) \quad (6)$$

Substituting our characteristic function $v(S) = -\text{Var}(\mathbb{E}[Y|x_S]) + a \cdot w(S)$ and rearranging:

$$\Delta_S - \Delta_T \geq a \cdot \Delta_w \quad (7)$$

where $\Delta_S = -\text{Var}(\mathbb{E}[Y|x_{S\cup\{i\}}]) + \text{Var}(\mathbb{E}[Y|x_S])$, $\Delta_T$ is defined similarly, and $\Delta_w = [w(T\cup\{i\}) - w(T)] - [w(S\cup\{i\}) - w(S)]$.

We prove that $\Delta_S - \Delta_T \leq 0$ by showing that for multivariate Gaussian distributions, the variance reduction from adding a new variable diminishes as the conditioning set grows. This follows from the information chain rule and can be expressed in terms of conditional mutual information:

$$\Delta_T - \Delta_S \propto I(Y; X_i|X_S) - I(Y; X_i|X_T) \leq 0 \quad (8)$$

When adding variable $i$ to the smaller set $S$, we get a larger variance reduction compared to adding $i$ to the larger set $T$, which means more supersets will newly satisfy the condition $v_{S'} < v_S$ than will newly satisfy $v_{S'} < v_T$. Therefore, the decrease in the penalty function is larger for $S$ (i.e., $w(S) - w(S\cup\{i\}) \geq w(T) - w(T\cup\{i\})$), which gives us $\Delta_w = [w(T \cup \{i\}) - w(T)] - [w(S \cup \{i\}) - w(S)] \geq 0$.

Therefore, the inequality $\Delta_S - \Delta_T \geq a \cdot \Delta_w$ is satisfied when $a \leq 0$ and $|a|$ is sufficiently large, guaranteeing the existence of a threshold $a^* \leq 0$ such that for all $a < a^*$, the characteristic function is convex. □

**Theorem 3.2** (Markov Boundary Attribution Property)**.** *Let $MB \subseteq X$ be the Markov boundary of $Y$ in the SCM. For any core solution $(\phi_1, \phi_2, \ldots, \phi_d)$ of the game with characteristic function $v(S) = -Var(\mathbb{E}[Y|x_S]) + a \cdot w(S)$ with appropriate $a$, the sum of attributions to variables in the Markov boundary exceeds the sum of attributions to variables outside the Markov boundary:*

$$\sum_{i \in MB} \phi_i \geq \sum_{j \in V \setminus \{Y \cup MB\}} \phi_j \quad (9)$$

The theorem shows that the method assign more importance to the features that are in a Markov boundary.

*Proof sketch of Theorem 3.2.* Let $MB \subseteq V \setminus \{Y\}$ be the Markov boundary of $Y$. By the Markov property, for

any superset $S$ where $MB \subseteq S$, the explained variance $\mathrm{Var}(\mathbb{E}[Y|X_S])$ equals the Markov Variance Explained (MVE). This implies $w(MB) = 0$.

For any core solution $(x_1, x_2, \ldots, x_d)$, the sum of attributions to variables in $MB$ satisfies $\sum_{i \in MB} x_i \geq v(MB) = MVE$. By efficiency, $\sum_{i \in V \setminus \{Y\}} x_i = v(V \setminus \{Y\}) = MVE$.

Let $N = V \setminus (\{Y\} \cup MB)$ be the variables outside the Markov boundary. Since $\sum_{i \in MB} x_i + \sum_{j \in N} x_j = MVE$ and $\sum_{i \in MB} x_i \geq MVE$, we have $\sum_{j \in N} x_j \leq 0$.

For any variable $k \in N$, the core property requires $\sum_{i \in MB} x_i + x_k \geq v(MB \cup \{k\}) = MVE$. Since $\sum_{i \in MB} x_i \geq MVE$, this implies $x_k \geq 0$. The only way to satisfy both constraints is if $\sum_{j \in N} x_j = 0$ and $x_k = 0$ for all $k \in N$. Therefore, all attributions go to the Markov boundary variables. □

## 4. Experimental demonstration

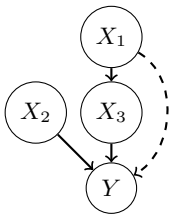


Figure 2: Directed graph representation. $X_1$– Gender, $X_2$– Test Score, $X_3$ – Department. In fair case, there is not direct link between $X_1$ and $Y$, while in the unfair case the edge $X_1 \to Y$ exists.

We evaluate our method using the college admission scenario from Frye et al. (2020). This setup involves predicting college admission based on three variables: gender, test score, and department. In the fair scenario, gender influences admission only indirectly through the number of applications to department (no direct causal path), while in the unfair scenario, gender has an effect on admission decisions, see Figure 2.

We implement the egalitarian least core solution with $L_2$ norm minimization (Benmerzoug & de Benito Delgado, 2023), incorporating a small numerical relaxation for computational stability. A small neural network was trained on synthetically generated data following the causal structure described in Frye et al. (2020). Similar to (Lundberg & Lee, 2017) we sampled dropped features from independent marginal distributions for simplicity.

Attributions of our new method are shown in Fig. 3. In the fair scenario, the attributions for gender, test score, and department are 0.182, 0.431, and 0.387 respectively, and 0.320, 0.363, and 0.316 in the unfair scenario. In the fair scenario, attribution to the Markov boundary variables (test score and department) exceeds the attribution to the non-Markov boundary variable (gender). In the unfair scenario, gender's attribution increases substantially from 0.182 to 0.320, reflecting its direct causal role. Notably, even in the fair case, gender is assigned non-zero attribution due to statistical imbalances in the data, correctly capturing its predictive utility.

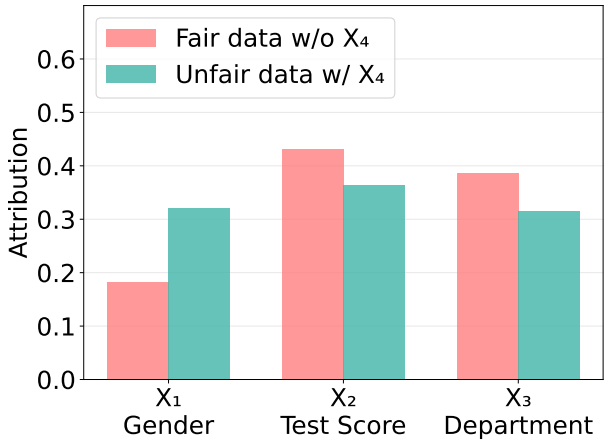


Figure 3: Attributions assigned by our method averaged across dataset.

## 5. Discussion and Future Work

**Applications to models** Our method demonstrates how to perform causal attribution in linear SCMs. We provided a small demonstration using a machine learning model in a simple case, assuming the model correctly learned the underlying distribution. However the connection between our theoretical framework and machine learning model predicting some values remains unclear and needs a further exploration and discussion. Additionally, although our theory is stated for linear models with additive Gaussian noise, the method can potentially be extended to nonlinear cases through piecewise linear approximation around local regions.

**Multiple Attribution Solutions** The core is a convex set rather than a single vector, allowing practitioners to choose any point within it. The Shapley value provides a natural default as it lies in the core for convex games (Madiman, 2008). Alternatively, practitioners can select core solutions that minimize attribution outside the Markov boundary, enforce sparsity, or minimize the $L_2$ norm via the egalitarian least-core (Benmerzoug & de Benito Delgado, 2023). This flexibility enables tailored attributions while preserving stability and causal-structure guarantees.

## References


Benmerzoug, A. and de Benito Delgado, M. [re] if you like shapley then you'll love the core. In *ML Reproducibility Challenge 2022*, 2023.

Frye, C., Rowat, C., and Feige, I. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in neural information processing systems*, 33:1229–1239, 2020.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33:4778–4789, 2020.

Janzing, D., Blöbaum, P., Mastakouri, A. A., Faller, P. M., Minorics, L., and Budhathoki, K. Quantifying intrinsic causal contributions via structure preserving interventions. In *International Conference on Artificial Intelligence and Statistics*, pp. 2188–2196. PMLR, 2024.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Ma, S. and Tourani, R. Predictive and causal implications of using shapley value for model interpretation. In *Proceedings of the 2020 KDD workshop on causal discovery*, pp. 23–38. PMLR, 2020.

Madiman, M. Cores of cooperative games in information theory. *EURASIP Journal on Wireless Communications and Networking*, 2008:1–12, 2008.

Pearl, J. *Causality*. Cambridge university press, 2009.

Shapley, L. S. Cores of convex games. *International journal of game theory*, 1:11–26, 1971.

Štrumbelj, E. and Kononenko, I. A general method for visualizing and explaining black-box regression models. In *Adaptive and Natural Computing Algorithms: 10th International Conference, ICANNGA 2011, Ljubljana, Slovenia, April 14-16, 2011, Proceedings, Part II 10*, pp. 21–30. Springer, 2011.


## A. Proofs

Proof of Variance Reduction Inequality in Multivariate Gaussian Distributions

**Theorem A.1.** *Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n, Y)$ be a multivariate Gaussian random vector with covariance matrix $\boldsymbol{\Sigma}$. For any sets $S \subseteq T \subseteq V \setminus \{Y\}$ and variable $i \in V \setminus \{Y \cup T\}$:*

$$[\mathrm{Var}(Y|S) - \mathrm{Var}(Y|S \cup \{i\})] \geq [\mathrm{Var}(Y|T) - \mathrm{Var}(Y|T \cup \{i\})] \tag{10}$$

*Proof.* We will approach this proof in multiple steps. First, we establish that the inequality stated in the theorem is equivalent to an inequality involving explained variances, then we prove the main result using properties of multivariate Gaussian distributions.

**Preliminary: Equivalence via Law of Total Variance**

We first show that for a multivariate Gaussian distribution, the following inequalities are equivalent:

$$[\mathrm{Var}(Y|S) - \mathrm{Var}(Y|S \cup \{i\})] \geq [\mathrm{Var}(Y|T) - \mathrm{Var}(Y|T \cup \{i\})] \tag{11}$$

$$[\mathrm{VarE}(Y|S \cup \{i\}) - \mathrm{VarE}(Y|S)] \geq [\mathrm{VarE}(Y|T \cup \{i\}) - \mathrm{VarE}(Y|T)] \tag{12}$$

where $\mathrm{VarE}(Y|S) = \mathrm{Var}(\mathbb{E}[Y|S])$ denotes the variance of the conditional expectation.

By the law of total variance:

$$\mathrm{Var}(Y) = E[\mathrm{Var}(Y|S)] + \mathrm{Var}(E[Y|S]) \tag{13}$$

For Gaussian distributions, the conditional variance $\mathrm{Var}(Y|S)$ is constant (does not depend on the specific values of $S$), so:

$$E[\mathrm{Var}(Y|S)] = \mathrm{Var}(Y|S) \tag{14}$$

Therefore:

$$\mathrm{Var}(Y|S) = \mathrm{Var}(Y) - \mathrm{Var}(E[Y|S]) \tag{15}$$

Similarly for $S \cup \{i\}$:

$$\mathrm{Var}(Y|S \cup \{i\}) = \mathrm{Var}(Y) - \mathrm{Var}(E[Y|S \cup \{i\}]) \tag{16}$$

Taking the difference:

$$\mathrm{Var}(Y|S) - \mathrm{Var}(Y|S \cup \{i\}) = \mathrm{Var}(E[Y|S \cup \{i\}]) - \mathrm{Var}(E[Y|S]) = \mathrm{VarE}(Y|S \cup \{i\}) - \mathrm{VarE}(Y|S) \tag{17}$$

The same relationship holds for set $T$. Since $\mathrm{VarE}(Y|S) = \mathrm{Var}(E[Y|S])$ for Gaussian distributions, the equivalence between inequalities equation 11 and equation 12 follows immediately.

**Prove the Main Inequality**

For the main inequality, we need to show:

$$[\text{Var}(Y|S) - \text{Var}(Y|S \cup \{i\})] \geq [\text{Var}(Y|T) - \text{Var}(Y|T \cup \{i\})] \tag{18}$$

Using our established relationship multivariate Gaussian distribution :

$$\text{Var}(Y|S) \cdot \rho^2_{Y,i|S} \geq \text{Var}(Y|T) \cdot \rho^2_{Y,i|T} \tag{19}$$

We need to establish two facts:

1. $\text{Var}(Y|S) \geq \text{Var}(Y|T)$ (since $S \subseteq T$)
2. $\rho^2_{Y,i|S} \geq \rho^2_{Y,i|T}$

The first fact follows directly from the properties of conditional variance: conditioning on more variables (or information) reduces variance.

For the second fact, we use information theory. For Gaussian random variables, the conditional mutual information can be expressed as:

$$I(Y; i|Z) = -\frac{1}{2}\log(1 - \rho^2_{Y,i|Z}) \tag{20}$$

A fundamental property of mutual information is that for $S \subseteq T$:

$$I(Y; i|S) \geq I(Y; i|T) \tag{21}$$

This is the data processing inequality: conditioning on more variables can only reduce the dependency between $Y$ and $i$.

From this inequality:

$$-\frac{1}{2}\log(1 - \rho^2_{Y,i|S}) \geq -\frac{1}{2}\log(1 - \rho^2_{Y,i|T}) \tag{22}$$

Since logarithm is a monotonically increasing function and the negative sign reverses the inequality:

$$\log(1 - \rho^2_{Y,i|S}) \leq \log(1 - \rho^2_{Y,i|T}) \tag{23}$$

Again, due to monotonicity of logarithm:

$$1 - \rho^2_{Y,i|S} \leq 1 - \rho^2_{Y,i|T} \tag{24}$$

Rearranging:

$$\rho^2_{Y,i|S} \geq \rho^2_{Y,i|T} \tag{25}$$

Now, combining our two established facts:

$$\text{Var}(Y|S) \geq \text{Var}(Y|T) \tag{26}$$

$$\rho^2_{Y,i|S} \geq \rho^2_{Y,i|T} \tag{27}$$

Therefore:

$$\text{Var}(Y|S) \cdot \rho^2_{Y,i|S} \geq \text{Var}(Y|T) \cdot \rho^2_{Y,i|T} \tag{28}$$

Which is equivalent to:

$$[\text{Var}(Y|S) - \text{Var}(Y|S \cup \{i\})] \geq \tag{29}$$

$$[\text{Var}(Y|T) - \text{Var}(Y|T \cup \{i\})] \tag{30}$$

This completes the proof, showing that the reduction in variance when adding variable $i$ decreases as we condition on more variables, which is a "diminishing returns" property of conditional variance in multivariate Gaussian distributions. □

### A.1. Proof that $\Delta_w \leq 0$

Setup and Definitions

Recall that the penalty function is defined as:

$$w(S) = |\{S' \subseteq V \backslash \{Y\} : S \subseteq S' \text{ and } \text{Var}(\mathbb{E}[Y|x_{S'}]) < \text{Var}(\mathbb{E}[Y|x_S])\}| \tag{31}$$

We need to prove that for $S \subseteq T \subseteq V \setminus \{Y\}$ and $i \in V \setminus \{Y \cup T\}$:

$$\Delta_w = [w(T \cup \{i\}) - w(T)] - [w(S \cup \{i\}) - w(S)] \leq 0 \tag{32}$$

For convenience, let us denote:

- $v_S = \text{Var}(\mathbb{E}[Y|x_S])$
- $v_{S\cup\{i\}} = \text{Var}(\mathbb{E}[Y|x_{S\cup\{i\}}])$
- $v_T = \text{Var}(\mathbb{E}[Y|x_T])$
- $v_{T\cup\{i\}} = \text{Var}(\mathbb{E}[Y|x_{T\cup\{i\}}])$

From the variance reduction property proven in the previous section, we know:

1. $v_S \geq v_{S\cup\{i\}}$ (adding variables reduces variance)
2. $v_T \geq v_{T\cup\{i\}}$ (adding variables reduces variance)
3. $v_S \geq v_T$ (since $S \subseteq T$)
4. $v_{S\cup\{i\}} \geq v_{T\cup\{i\}}$ (since $S \cup \{i\} \subseteq T \cup \{i\}$)

#### A.1.1. Key Observation

First, observe that both $w(S \cup \{i\}) - w(S)$ and $w(T \cup \{i\}) - w(T)$ are non-positive. This is because:

- A superset $S'$ is counted in $w(S)$ if $S \subseteq S'$ and $v_{S'} < v_S$

- A superset $S'$ is counted in $w(S\cup\{i\})$ if $(S\cup\{i\}) \subseteq S'$ and $v_{S'} < v_{S\cup\{i\}}$

Since $(S \cup \{i\}) \subseteq S'$ implies $S \subseteq S'$, and $v_{S\cup\{i\}} \leq v_S$, every superset counted in $w(S \cup \{i\})$ is also counted in $w(S)$. Therefore, $w(S \cup \{i\}) \leq w(S)$.

### A.1.2. Categorization of Supersets

Let us categorize supersets $S'$ based on whether they contain $i$:

**Type 1: Supersets without $i$ (i.e., $i \notin S'$)**

- These contribute to $w(S)$ if $S \subseteq S'$ and $v_{S'} < v_S$
- These contribute to $w(T)$ if $T \subseteq S'$ and $v_{S'} < v_T$
- These never contribute to $w(S \cup \{i\})$ or $w(T \cup \{i\})$ since they do not contain $i$

**Type 2: Supersets with $i$ (i.e., $i \in S'$)**

- These contribute to $w(S)$ if $S \subseteq S'$ and $v_{S'} < v_S$
- These contribute to $w(S \cup \{i\})$ if $S \subseteq S'$ (automatic since $i \in S'$) and $v_{S'} < v_{S\cup\{i\}}$
- These contribute to $w(T)$ if $T \subseteq S'$ and $v_{S'} < v_T$
- These contribute to $w(T \cup \{i\})$ if $T \subseteq S'$ and $v_{S'} < v_{T\cup\{i\}}$

### A.1.3. Computing the Differences

The difference $w(S) - w(S \cup \{i\})$ counts:

1. All Type 1 supersets that satisfy $S \subseteq S'$ and $v_{S'} < v_S$
2. Type 2 supersets where $v_{S\cup\{i\}} \leq v_{S'} < v_S$

Similarly, $w(T) - w(T \cup \{i\})$ counts:

1. All Type 1 supersets that satisfy $T \subseteq S'$ and $v_{S'} < v_T$
2. Type 2 supersets where $v_{T\cup\{i\}} \leq v_{S'} < v_T$

### A.1.4. Establishing the Inequality

To prove $\Delta_w \leq 0$, we need to show:

$$w(T) - w(T \cup \{i\}) \geq w(S) - w(S \cup \{i\}) \tag{33}$$

Let us denote:

- $\mathcal{A}_S$ = Type 1 supersets contributing to $w(S)$ but not $w(S \cup \{i\})$
- $\mathcal{A}_T$ = Type 1 supersets contributing to $w(T)$ but not $w(T \cup \{i\})$
- $\mathcal{B}_S$ = Type 2 supersets with $v_{S\cup\{i\}} \leq v_{S'} < v_S$
- $\mathcal{B}_T$ = Type 2 supersets with $v_{T\cup\{i\}} \leq v_{S'} < v_T$

Then:

$$w(S) - w(S \cup \{i\}) = |\mathcal{A}_S| + |\mathcal{B}_S| \tag{34}$$

$$w(T) - w(T \cup \{i\}) = |\mathcal{A}_T| + |\mathcal{B}_T| \tag{35}$$

$\mathcal{A}_T \subseteq \mathcal{A}_S$

*Proof.* If $S' \in \mathcal{A}_T$, then $T \subseteq S'$, $i \notin S'$, and $v_{S'} < v_T$. Since $S \subseteq T \subseteq S'$, we have $S \subseteq S'$. Since $v_S \geq v_T > v_{S'}$, we have $v_{S'} < v_S$. Therefore, $S' \in \mathcal{A}_S$. □

$\mathcal{B}_T \subseteq \mathcal{B}_S$

*Proof.* From the variance reduction inequality proven in the previous section, we know that:

$$v_S - v_{S\cup\{i\}} \geq v_T - v_{T\cup\{i\}} \tag{36}$$

This means the interval $[v_{S\cup\{i\}}, v_S)$ contains the interval $[v_{T\cup\{i\}}, v_T)$.

If $S' \in \mathcal{B}_T$, then:

- $T \subseteq S'$ (which implies $S \subseteq S'$ since $S \subseteq T$)
- $i \in S'$
- $v_{T\cup\{i\}} \leq v_{S'} < v_T$

Since $[v_{T\cup\{i\}}, v_T) \subseteq [v_{S\cup\{i\}}, v_S)$, we have $v_{S\cup\{i\}} \leq v_{S'} < v_S$. Therefore, $S' \in \mathcal{B}_S$. □

### A.1.5. Conclusion

From Claims 1 and 2:

- $\mathcal{A}_T \subseteq \mathcal{A}_S$ implies $|\mathcal{A}_T| \leq |\mathcal{A}_S|$
- $\mathcal{B}_T \subseteq \mathcal{B}_S$ implies $|\mathcal{B}_T| \leq |\mathcal{B}_S|$

Therefore:

$$w(T) - w(T \cup \{i\}) = |\mathcal{A}_T| + |\mathcal{B}_T| \leq |\mathcal{A}_S| + |\mathcal{B}_S| = w(S) - w(S \cup \{i\}) \tag{37}$$

To connect this to $\Delta_w$, recall that:

$$\Delta_w = [w(T \cup \{i\}) - w(T)] - [w(S \cup \{i\}) - w(S)] \tag{38}$$

We can rewrite this as:

$$\Delta_w = -[w(T) - w(T \cup \{i\})] - (-[w(S) - w(S \cup \{i\})]) \tag{39}$$

Since $w(T) - w(T \cup \{i\}) \leq w(S) - w(S \cup \{i\})$, multiplying by $-1$ gives:

$$-[w(T) - w(T \cup \{i\})] \geq -[w(S) - w(S \cup \{i\})] \tag{40}$$

Therefore:

$$\Delta_w = -[w(T) - w(T \cup \{i\})] + [w(S) - w(S \cup \{i\})] \tag{41}$$

$$\geq -[w(S) - w(S \cup \{i\})] + [w(S) - w(S \cup \{i\})] \tag{42}$$

$$= 0 \tag{43}$$

### A.2. Detailed Proof of Theorem 2: Markov boundary Attribution Property

*Proof.* Let $MB \subseteq V \setminus \{Y\}$ be the Markov boundary of $Y$ in the SCM. Define the Markov Variance Explained (MVE) as:

$$MVE = \text{Var}(\mathbb{E}[Y|X_{MB}]) \tag{44}$$

We first establish a key property: For any super set $S$ such that $MB \subseteq S \subseteq V \setminus \{Y\}$, the explained variance remains constant:

$$\text{Var}(\mathbb{E}[Y|X_S]) = MVE \tag{45}$$

This follows directly from the definition of the Markov boundary, which contains all the information necessary to predict $Y$. Adding variables outside the Markov boundary does not improve predictive power. As a consequence, $w(S) = 0$ for all such super sets, including $w(MB) = 0$ and $w(V \setminus \{Y\}) = 0$.

Now, consider any core solution $(x_1, x_2, \ldots, x_d)$ of our game. By the definition of the core, for any coalition $S$:

$$\sum_{i \in S} x_i \geq v(S) \tag{46}$$

Applying this to the Markov boundary:

$$\sum_{i \in MB} x_i \geq v(MB) = MVE \tag{47}$$

By the efficiency property of core solutions:

$$\sum_{i \in V \setminus \{Y\}} x_i = v(V \setminus \{Y\}) = MVE \tag{48}$$

Letting $N = V \setminus (\{Y\} \cup MB)$ represent variables outside the Markov boundary, we have:

$$\sum_{i \in MB} x_i + \sum_{j \in N} x_j = MVE \tag{49}$$

From our earlier inequality $\sum_{i \in MB} x_i \geq MVE$, we must have:

$$\sum_{j \in N} x_j \leq 0 \tag{50}$$

For any individual variable $k \in N$, consider the coalition $MB \cup \{k\}$. By the Markov boundary property:

$$v(MB \cup \{k\}) = \text{Var}(\mathbb{E}[Y|X_{MB \cup \{k\}}]) + a \cdot w(MB \cup \{k\}) = MVE \tag{51}$$

The core property requires:

$$\sum_{i \in MB} x_i + x_k \geq v(MB \cup \{k\}) = MVE \tag{52}$$

Since we already know $\sum_{i \in MB} x_i \geq MVE$, this implies $x_k \geq 0$.

Now we have two constraints:

$$\sum_{j \in N} x_j \leq 0 \tag{53}$$

$$x_k \geq 0 \quad \forall k \in N \tag{54}$$

The only way both can be satisfied simultaneously is if $\sum_{j \in N} x_j = 0$ and $x_k = 0$ for all $k \in N$.

This means that variables outside the Markov boundary receive zero attribution, while variables in the Markov boundary receive all the attribution:

$$\sum_{i \in MB} x_i = MVE > 0 = \sum_{j \in V \setminus \{Y \cup MB\}} x_j \tag{55}$$

This completes the proof, establishing that our attribution method correctly identifies the Markov boundary as the most important set of variables for predicting $Y$. □