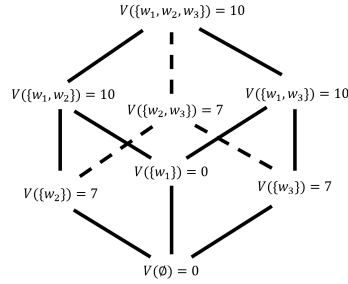


A APPENDIX

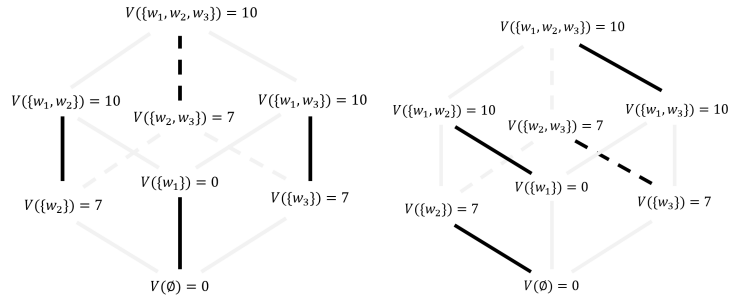
An Example of SV-Based Pruning Failure. Three parameters, denoted by w_1, w_2, w_3 , produce a total gain of 10 when all are used. Let them have an f^* -hypercube with the following vertex values:

$$f^*(S) = \begin{cases} 10, & \text{if } S = \{w_1, w_2, w_3\} \\ 10, & \text{if } S = \{w_1, w_2\}, \\ 10, & \text{if } S = \{w_1, w_3\}, \\ 7, & \text{if } S = \{w_2, w_3\}, \\ 7, & \text{if } S = \{w_2\}, \\ 7, & \text{if } S = \{w_3\}, \\ 0, & \text{if } S = \{w_1\}, \\ 0, & \text{if } S = \emptyset. \end{cases} \quad (1)$$

The corresponding f^* values represent the incremental gains associated with transitions along vertex paths. The SVs assigned to each parameter are $(w_1, w_2, w_3) = (2, 4, 4)$, indicating that w_1 should be pruned first. On the other hand, CI yields $(w_1, w_2, w_3) = (\frac{2}{3}, \frac{1}{2}, \frac{1}{2})$ reflecting the redundancy among parameters. According to the CI, w_2 or w_3 should be pruned, as they are replaceable by each other, preserving the total gain after pruning.



(a) geometric view of all possible path



(b) All possible path of w_1

(c) All possible path of w_2

Figure 1: Visualization of the f^* -hypercube in Appendix A.

B APPENDIX B

The Impact of Parameter Removal on SHAP Value. Let \mathcal{P} be the set of parameters, and let $f(S)$ be a performance function with $S \in 2^{\mathcal{P}}$. Assume that the following Non-negative Contribution and Replaceability conditions hold for parameter $k \in \mathcal{P}$:

1. **Non-negative Contribution:** The marginal contribution after adding a new parameter $k \notin S$ does not decrease. For f representing the optimized performance in this paper, this condition is automatically satisfied.

$$\Delta(k|S) \equiv f(S \cup \{k\}) - f(S) \geq 0.$$

2. **Replaceability:** The marginal contribution of k is zero in the presence of any parameter $l \notin S$:

$$f(S \cup \{l\} \cup \{k\}) - f(S \cup \{l\}) = 0.$$

Then the following hold

$$\Delta(l|S \cup \{k\}) \leq \Delta(l|S), \quad (2)$$

for all $l \notin S \cup \{k\}$ with $S \subset \mathcal{P} \setminus \{k\}$. In other words, pruning of k increases the marginal contribution of other parameters.

Proof.

$$\begin{aligned} \Delta(l|S \cup \{k\}) &= f(\{l\} \cup S \cup \{k\}) - f(S \cup \{k\}) \\ &= f(\{l\} \cup S \cup \{k\}) - f(S \cup \{k\}) + \{f(S) - f(S)\} \\ &\leq f(\{l\} \cup S \cup \{k\}) - f(S) \quad (\because f(S \cup \{k\}) - f(S) \geq 0, \text{Assumption1}) \\ &= f(\{l\} \cup S \cup \{k\}) - f(S) + \{f(\{l\} \cup S) - f(\{l\} \cup S)\} \\ &= f(\{l\} \cup S) - f(S) \quad (\because f(\{l\} \cup S \cup \{k\}) - f(\{l\} \cup S) = 0, \text{Assumption2}) \\ &= \Delta(l|S). \end{aligned}$$

□

Performance Retention under Pruning. With subsets $T \subseteq \mathcal{P} \setminus \{l\}$ and $S \subseteq \mathcal{P} \setminus \{l, k\}$, the SV of $l \in \mathcal{P}$ can be written as

$$\phi_l = \sum_{T \subseteq \mathcal{P} \setminus \{l\}} \underbrace{\frac{|T|! (n - |T| - 1)!}{n!}}_{w_T} \Delta(l | T) \quad (3)$$

$$= \sum_{S \subseteq \mathcal{P} \setminus \{l, k\}} \left[\underbrace{\frac{|S|! (n - |S| - 1)!}{n!}}_{w_S} \Delta(l | S) + \underbrace{\frac{|S+1|! (n - |S| - 2)!}{n!}}_{w_{S \cup \{k\}}} \Delta(l | S \cup \{k\}) \right]. \quad (4)$$

Let's define

$$w'_S \equiv w_S + w_{S \cup \{k\}} = \frac{|S|! (n - |S| - 2)!}{(n - 1)!}. \quad (5)$$

Then the SV for l is always less than the SV for l after pruning k .

$$\phi_l = \sum_{S \subseteq \mathcal{P} \setminus \{l, k\}} [w_S \Delta(l | S) + w_{S \cup \{k\}} \Delta(l | S \cup \{k\})] \quad (6)$$

$$\leq \sum_{S \subseteq \mathcal{P} \setminus \{l, k\}} (w_S + w_{S \cup \{k\}}) \Delta(l | S) \quad (\text{Eq. 2}) \quad (7)$$

$$= \sum_{S \subseteq \mathcal{P} \setminus \{l, k\}} w'_S \Delta(l | S) \quad (8)$$

$$= \phi_l^{\sim k}, \quad (9)$$

where

$$\phi_l^{\sim k} = \sum_{S \subseteq \mathcal{P} \setminus \{l, k\}} \underbrace{\frac{|S|! (n - |S| - 2)!}{(n - 1)!}}_{w'_S} \Delta(l \mid S).$$

Summing this inequality over all remaining parameters yields

$$J_{\text{tot}}(\mathcal{P} \setminus \{k\}) = \sum_{l \in \mathcal{P} \setminus \{k\}} \phi_l^{\sim k} \geq \sum_{l \in \mathcal{P} \setminus \{k\}} \phi_l = J_{\text{tot}}(\mathcal{P}) - \phi_k. \quad (10)$$

Therefore, under the Non-negativity contribution and replacability conditions, performance degradation by removing parameter k is compensated by other remaining parameters and always less than k 's SV.

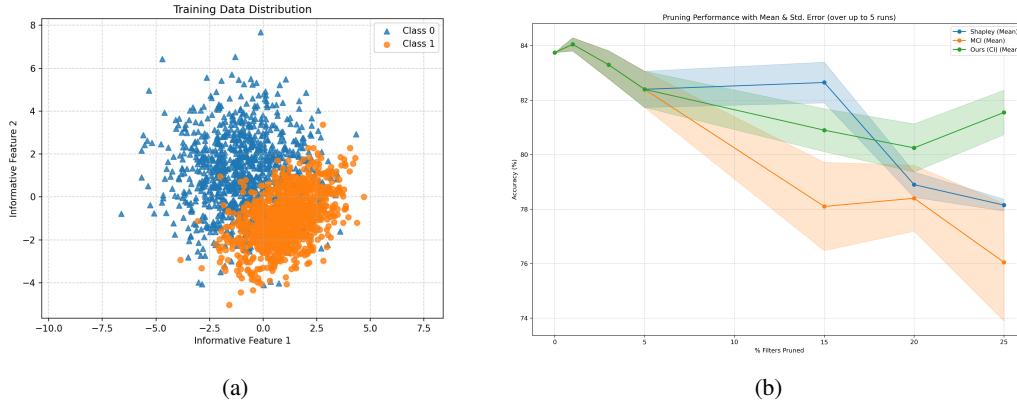


Figure 2: (a) Visualization of 100-dimensional training data on the 2-D plane. (b) Pruning results for SV, MCI and CI.

C APPENDIX C

Synthetic Experiment. We use two 100-dimensional Gaussians, where the first two dimensions are shown in Figure 2a and the remaining 99 dimensions are filled with standard Gaussian noise. We use a fully connected neural network with three layers, each containing 16 hidden units. The data is designed to be easily classified so that the pruned neural network possesses only enough capacity to identify simple decision boundaries. The two-level approximation scheme is used. We sample 1,000 permutations and 100 vertices for a regression function. For each vertex sample, the corresponding model is trained for 10 epochs with learning rate (0.01). For f^* , the optimized cross entropy loss is used.

Results. The pruning results are presented in Figure 2b. Here, we report the mean and standard error of test accuracy from 5 independent simulations. The pruning ratio is defined as the ratio of the number of pruned parameters to the total number of parameters. While all three methods perform reasonably well at low pruning ratios ($\leq 5\%$), the MCI methods extremely fail to capture the discriminative structure after 5%. Thus, the MCI method is excluded from baselines of real-world experiments.