If You Like Shapley Then You’ll Love the Core

Tom Yan ¹  Ariel D. Procaccia ²

Abstract

The prevalent approach to problems of credit assignment in machine learning — such as feature attribution and data valuation — is to model the problem at hand as a cooperative game and solve it using the Shapley value. But cooperative game theory offers a rich menu of alternative solution concepts, which famously includes the core and its variants. Our goal is to challenge the machine learning community’s surprising consensus around the Shapley value, and make a case for the core as a viable alternative. To that end, we prove that arbitrarily good approximations to the least core — a core relaxation that is always feasible — can be computed efficiently (but prove an impossibility for a more refined solution concept, the nucleolus). We also perform experiments that corroborate these theoretical results.

1. Introduction

As machine learning systems become more sophisticated and capable, they are increasingly used in our society to automate tasks and generate value. This has led to a surge in the attention given to explainability for machine learning: how features and data contribute to the performance of ML models. To ensure ML models are functioning as intended, much work has been devoted to studying feature attribution: how the features used to represent the data influence the model’s predictions (Cohen et al., 2007; Štrumbelj & Kononenko, 2010; Datta et al., 2015; 2016; Lundberg & Lee, 2017; Chen et al., 2019). Related to feature attribution is data valuation (Ghorbani & Zou, 2019; Jia et al., 2019a;b; Ohrimenko et al., 2019; Agarwal et al., 2019), which studies how data points contribute to model performance. With ML models now generating profit for enterprises, this understanding is important in order to fairly compensate data suppliers for their training data. Central to both pursuits is an equitable means of credit assignment.

Virtually all papers, including every single paper cited above, deem the Shapley value (or close variants thereof) to be the “right” way to carry out credit assignment. The Shapley value is a solution concept from cooperative game theory in which players — in this case features or data points — are assigned payoffs in a way that satisfies four axioms; roughly speaking, a player’s payoff is their average marginal contribution to a coalition consisting of other players.

This intense focus on the Shapley value is rather baffling, however, as — once we have accepted that problems involving credit assignment in machine learning can be modeled as cooperative games — there are a plethora of other solution concepts (Peleg & Sudhölter, 2007). In particular, there is a seminal solution concept in cooperative game theory that is viewed as being as prominent as the Shapley value: the core. This solution concept seeks to achieve maximal stability amongst all possible coalitions of the players in the game — an idea that dates back to the writings of Edgeworth in 1881. Since then, it has found extensive applications in economics and beyond (Telser, 1994).

Studies in behavioral game theory have found the core to be predictive of payment distribution in market settings, suggesting that people perceive the core as a fair scheme for dividing up the total payoff; by contrast, the Shapley value has received “weaker empirical support” (Williams, 1988). This is an especially compelling reason to prefer the core over the Shapley value: since the stakeholders involved with machine learning are often people, it is imperative to employ a solution concept that is consistent with their behavior and intuition (Bhatt et al., 2019).

¹School of Computer Science, Carnegie Mellon University
²School of Engineering and Applied Sciences, Harvard University.
Correspondence to: Tom Yan <tyan2@andrew.cmu.edu>, Ariel Procaccia <arielpro@seas.harvard.edu>.
In this paper, we aim to show that the (least) core is a computationally viable alternative to the Shapley value for credit assignment in machine learning. In doing so, we hope to raise awareness of the core as a natural solution concept for fair credit assignment, challenge the wide-ranging usage of the Shapley value and inspire a closer examination of cases where one solution concept should be preferred over the other.

1.1. Our Results

Much like the Shapley value, the primary obstacle in applying the concept of least core is computational complexity. Indeed, it is the solution to a linear program whose number of constraints is exponential in the number of players. Nevertheless, we construct a Monte Carlo algorithm that runs in polynomial time and (with given confidence) outputs a payoff allocation in the $\delta$-probable least core — a slightly relaxed version of the least core where the payoff constraints may be violated by up to a $\delta$-fraction of coalitions. When the number of players is large, though, this may still be intractable; we therefore show that it is possible to find a solution in the $(\epsilon, \delta)$-probably approximate least core — whose constraints are additionally relaxed by $\epsilon$ each — in time that is polylogarithmic in the number of players.

We also study a well-known and compelling refinement of the least core called the nucleolus. However, it turns out that results that are analogous to those for the least core are essentially unattainable. Informally, we prove that any algorithm would require access to the values of an exponentially large number of coalitions to compute a payoff allocation in the $(\epsilon, \delta)$-probably approximate nucleolus, which again relaxes all relevant constraints by $\epsilon$ and allows a $\delta$-fraction of the constraints to be violated. The juxtaposition of the positive computational results for the least core and the negative result for the nucleolus provides a strong endorsement of the former (somewhat coarser) notion over the latter.

To verify these results empirically, we compute the $\delta$-probable least core and the $(\epsilon, \delta)$-probably approximate nucleolus for feature attribution tasks, using three classification datasets that originate from different domains. The experimental results confirm our theoretical finding that it is computationally tractable to approximate the least core, but not the nucleolus. We also conduct experiments comparing the least core with the Shapley value; we find that (as expected) the least core leads to significantly more stable payoff allocations in practice, and (surprisingly) the least core can be seen as being more discriminating. Finally, in a data valuation setup, our results suggest that the least core is a good measure of the influence of data on a classifier’s accuracy.

1.2. Related Work

There is an entire area of algorithmic game theory devoted to the computation of solutions of cooperative games (Chalkiadakis et al., 2011). In particular, a slew of papers have studied the complexity of the core, the least core, and the nucleolus in specific classes of cooperative games (Deng & Papadimitriou, 1994; Conitzer & Sandholm, 2006; Bachrach & Rosenschein, 2008; Elkind & Pasechnik, 2009; Elkind et al., 2009).

Our work is most closely related to that of Balkanski et al. (2017). They study settings where solutions to cooperative games — specifically, the Shapley value and the core — are learned from samples consisting of coalitions and their values. Like Balcan et al. (2015), they are motivated by the observation that in classical applications of cooperative games values of coalitions cannot be accessed via queries; for example, if the game represents company employees working together to complete tasks, it is impossible to know which tasks would be completed had a specific coalition worked alone. Under the assumption that the underlying game has a nonempty core, Balkanski et al. give bounds on the sample complexity of three approximations of the core.

On a technical level, our positive results for the least core (Theorems 1 and 2) extend the corresponding definitions and results of Balkanski et al. for the core, by eschewing the assumption that the core is nonempty; our proofs of these results directly build on theirs. Our interpretation of these results is quite different, though, because in our setting coalition values can be queried — for example, one can run a black-box predictor with a specific subset of features and measure its accuracy — so we think of our results as guarantees on the performance of Monte Carlo algorithms. Balkanski et al. did not study the nucleolus, so our negative result for the nucleolus (Theorem 3) — which we view as our main theoretical result — has no analog in their work. Finally, the work of Balkanski et al. is purely theoretical, whereas our empirical results demonstrate the applicability of the least core to credit assignment in machine learning.

2. Preliminaries

A cooperative game consists of a set of players $N = \{1, \ldots, n\}$ and a characteristic function $v : 2^N \to \mathbb{R}$ which assigns a value to each coalition $S \subseteq N$, such that $v(\emptyset) = 0$; we assume that $v(S) \geq 0$ and $v(S) \leq 1$ for all $S \subseteq N$ for ease of exposition. We think of $v(S)$ as the payoff the coalition $S$ could obtain if it went it alone. Given such a game, we are interested in finding a payoff allocation (also known as an imputation) $x = (x_1, \ldots, x_n)$, where $x_i$ is the payoff of player $i \in N$. The payoff allocation must be efficient, that is,

$$\sum_{i \in N} x_i = v(N).$$
A payoff allocation is in the $e$-core if and only if the total payoff of each coalition is at least its value, up to $e$:
\[ \forall S \subseteq N, \sum_{i \in S} x_i + e \geq v(S). \]

The core itself, by this definition, satisfies these constraints with $e = 0$. Unfortunately, there are cooperative games whose core is empty (see Example 1 below). But clearly the $e$-core is nonempty if $e$ is large enough.

The idea behind the least core (Maschler et al., 1979) is to choose the smallest $e$ possible. It may be defined as the set of all solutions to the following linear program.

\[
\begin{align*}
\min & \quad e \\
\text{s.t.} & \quad \sum_{i \in N} x_i = v(N) \\
& \quad \sum_{i \in S} x_i + e \geq v(S) \quad \forall S \subseteq N
\end{align*}
\]

One can think of the least core as the set of payoff allocations that require the smallest subsidy $e^*$ (the value of $e$ in the optimal solution to (1)) to each coalition so that, if the payoff to each coalition was boosted by $e^*$, the allocation would be in the core. The core is nonempty if and only if $e^* \leq 0$.

We next consider a refinement of the least core, the nucleolus, first proposed by Schmeidler (1969). Define the deficit of a payoff allocation $x$ for a coalition $S \subseteq N$ to be $v(S) - \sum_{i \in S} x_i$. The nucleolus is the payoff allocation whose sorted list of deficits across all coalitions lexicographically dominates the list of deficits for any other payoff allocation. That is, the largest deficit (which will be positive if the core is empty) should be as small as possible; subject to that, the second largest deficit should be as small as possible, and so on. Notice that, in particular, the nucleolus minimizes the largest deficit, which leads to an allocation in the least core. In contrast to the least core, which may contain multiple payoff allocations, the nucleolus is known to be unique (Schmeidler, 1969).

**Example 1.** Consider a cooperative game with $N = \{1, 2, 3\}$ and the following characteristic function:
\[ v(\{1\}) = 0, v(\{2\}) = 0.7, v(\{3\}) = 0.4, v(\{1, 2\}) = 0.6, v(\{1, 3\}) = 0.2, v(\{2, 3\}) = 0.9, v(\{1, 2, 3\}) = 0.6. \]

The core is clearly empty because $v(N) < v(\{2, 3\})$. There are multiple least core allocations with $e^* = 0.3$, including $x^1 = (0, 0.45, 0.15)$ and $x^2 = (0, 0.4, 0.2)$.

The sorted list of deficits of the payoff allocation $x^1$ is $(0.3, 0.25, 0.25, 0.15, 0.05, 0, 0)$, and the sorted list of deficits of $x^2$ is $(0.3, 0.3, 0.2, 0.2, 0, 0)$. The former lexicographically dominates the latter because the second coordinate is smaller. Indeed, $x^1$ is the nucleolus.

### 3. Theoretical Results

Exact computation of the least core and the nucleolus requires solving linear programs with as many constraints as there are coalitions, which would typically be prohibitively expensive. Our strategy, therefore, is to sample a relatively small number of coalitions from an underlying distribution, and compute the desired solution concept on the sampled coalitions — this can be done in time that is polynomial in the number of samples, via the linear program (1) for the least core, and via a sequence of such linear programs for the nucleolus (Kopelowitz, 1967). The hope is that this Monte Carlo algorithm would give us a payoff allocation that approximates the desired one with respect to the underlying distribution.

#### 3.1. Computing the Least Core

We know from the work of Balkanski et al. (2017) that computing the least core exactly is a nonstarter — they prove an impossibility even for the core, under the assumption that it is nonempty. We therefore consider approximate versions of the least core.

Given a cooperative game, let $D$ be a distribution over $2^N$, and let $e^*$ be the subsidy defined by the least core — the optimal solution to Equation (1). A payoff allocation $x$ is in the $\delta$-probable least core if and only if
\[ \Pr_{S \sim D} \left( \sum_{i \in S} x_i + e^* \geq v(S) \right) \geq 1 - \delta. \]

That is, the least core constraint is violated with probability at most $\delta$ when coalitions are drawn from $D$.

We have the following result, whose proof appears in Appendix A.

**Theorem 1.** Given a cooperative game $(N,v)$, distribution $D$ over $2^N$, and $\delta, \Delta > 0$, solving the linear program (1) over $O((n + \log(1/\Delta))/\delta^2)$ coalitions sampled from $D$ gives a payoff allocation in the $\delta$-probable least core with probability at least $1 - \Delta$.

Note that the choice of $D$ rests with the algorithm designer. In other words, we can sample coalitions from any distribution $D$ and compute an allocation in the least core on the sample; the probable least core guarantee would then hold with respect to that same $D$. In particular, if the uniform distribution over coalitions is used, the guarantee holds with respect to a $(1 - \delta)$-fraction of all coalitions.

While Theorem 1 is encouraging, a potential drawback is that the algorithm’s running time is polynomial in the number of players $n$. While this is an exponential improvement over naïve least core computation, it can still be a nonstarter when the players are features in a high-dimensional space or data points. We therefore define the $(\epsilon, \delta)$-probably approximate least core to be payoff allocations such that
\[ \Pr_{S \sim D} \left( \sum_{i \in S} x_i + e^* + \epsilon \geq v(S) \right) \geq 1 - \delta. \]
We also note that the theorem statement deals with algorithms that are deterministic, up to the random sampling of coalitions from $\mathcal{D}$. However, it is not difficult to extend the theorem to deal with randomized algorithms too, at the cost of complicating the proof further. In addition, the constants in the theorem statement can certainly be improved, but we do not view their exact values as being important.

The rest of this section is devoted to the proof of Theorem 3. On a high level, we will construct a set of cooperative games $\mathcal{G}$ over the same set of players $N$, and a distribution $\mathcal{D}$ over the coalitions, such that no deterministic algorithm can compute a payoff allocation in the $(\epsilon, \delta)$-approximately least core with probability at least $1 - \Delta$. Intuitively, then, when such an input is observed, the algorithm does not have enough information about the underlying game and is likely to violate the $(\epsilon, \delta)$-approximately nucleolus requirement. In the theorem’s proof itself, we formalize this intuition by first assuming that the game itself is drawn from a uniform distribution over $\mathcal{G}$; the theorem statement follows from an averaging argument.

Formally, the class of games $\mathcal{G}$ is defined as follows. Let $N$ be a set of $n$ players; we assume without loss of generality that $n$ is divisible by 3. Let $C_1$ be a set of 3 players $\{i, j, k\}$. Define $C_2, C_3, C_4$ to be sets of $n/3 - 1$ players such that $C_1 \cup C_2 \cup C_3 \cup C_4 = N$. Each cooperative game $G_{C_1, C_2, C_3, C_4}$ in our class $\mathcal{G}$ is such that $v(S) = 1$ if $\{i, j\} \cup C_2 \subseteq S$ or $\{i, k\} \cup C_3 \subseteq S$ or $\{j, k\} \cup C_4 \subseteq S$; $v(S) = 0$ otherwise. The important thing to note is that all coalitions of size $n/3 + 1$ have value 0, except for exactly three that have value 1: $\{i, j\} \cup C_2$, $\{i, k\} \cup C_3$, and $\{j, k\} \cup C_4$. We call $C_1$ the critical set of game $G_{C_1, C_2, C_3, C_4}$.

Next, we define the distribution $\mathcal{D}$ to be the uniform distribution over all coalitions of size $n/3 + 1$.

**Lemma 1.** For any $m$ coalitions $S_1, \ldots, S_m$ of size $n/3 + 1$, at least half of the games in $\mathcal{G}$ satisfy $v(S_i) = 0$ for all $i = 1, \ldots, m$.

**Proof.** To count the number of such games, we can count the number of games in which the value of $S_i$ is 1. By symmetry, the number of games in which a coalition $S$ has value 1 is the same for all coalitions $S$ of size $n/3 + 1$. Moreover, for each game in $\mathcal{G}$ there are three coalitions of size $n/3 + 1$ with value 1. Therefore, for each $S_i$, the number of games in $\mathcal{G}$ with $v(S_i) = 1$ is $3|\mathcal{G}|/\binom{n}{n/3+1}$. It follows that the number of games for which it does not hold that $v(S_i) = 0$ for all $i = 1, \ldots, m$ is at most $3m|\mathcal{G}|/\binom{n}{n/3+1}$.
We claim that this payoff allocation is the only one that achieves the probably approximate nucleolus for games from at most one equivalence class. We know from Lemma 2 that all but one of these coalitions have value 0 and deficit $-2/3$. In order for the property

$$
\sum_{i \in S} x_i + d^*(S) - v(S) \leq \epsilon
$$

to hold for such coalitions, we would need their payoff to be at least $2/3 - \epsilon$.

Overall, there are at least $\binom{n-3}{n/3-1} - \delta \binom{n}{n/3+1} - 1$ many coalitions containing $i,j$ but not $k$ for which Equation (2) applies and have value 0. The middle term comes from factoring in that at most a $\delta$ fraction of all $\binom{n}{n/3+1}$ coalitions will not satisfy the probably approximate nucleolus property. By summing over the total payoffs of all such coalitions we have

$$
\left( \frac{n-3}{n/3-1} \right) (x_i + x_j) + \left( \frac{n-4}{n/3-2} \right) \left( \sum_{t \notin \{i,j,k\}} x_t \right) 
\geq \left( \frac{n-3}{n/3-1} - \delta \right) \left( \frac{n}{n/3+1} - 1 \right) (2/3 - \epsilon)
$$

since each player that is not $i,j$ or $k$ shows up $\binom{n-4}{n/3-2}$ times. Dividing by $\binom{n-3}{n/3-1}$ and using the fact that $\binom{n-4}{n/3-2}/\binom{n-3}{n/3-1} = 1/3$, we have

$$
x_i + x_j + 1/3 \left( \sum_{t \notin \{i,j,k\}} x_t \right) \geq 1 - \frac{n}{(n/3+1) \cdot \delta - \frac{1}{n/3-1}} (2/3 - \epsilon).
$$
With $n \geq 9$, $$\frac{1}{(n-3)} \leq \frac{1}{15}$$ and so we obtain

$$x_i + x_j + \frac{1}{3} \left( \sum_{v \notin \{i,j,k\}} x_t \right) \geq \left( \frac{14}{15} - \frac{n(n-1)(n-2)}{(n-3)(2n-3)} \delta \right) \left( \frac{2}{3} - \epsilon \right).$$

Using efficiency, $$\sum_{t \notin \{i,j,k\}} x_t = 1 - x_i - x_j - x_k,$$ and using the fact that $$\frac{(\frac{n}{3}+1)}{(\frac{n}{3}-1)} = \frac{n(n-1)(n-2)}{(n-3)(2n-3)} \leq 27$$ we get

$$\frac{2}{3} x_i + \frac{2}{3} x_j + \frac{1}{3} - \frac{1}{3} x_k \geq \left( \frac{14}{15} - 27\delta \right) \left( \frac{2}{3} - \epsilon \right).$$

Similarly, by considering the set of all coalitions that contain $i, k$ but not $j$, we see that

$$\frac{2}{3} x_i + \frac{2}{3} x_k + \frac{1}{3} - \frac{1}{3} x_j \geq \left( \frac{14}{15} - 27\delta \right) \left( \frac{2}{3} - \epsilon \right).$$

Summing both inequalities, we conclude that

$$\frac{4}{3} x_i + \frac{1}{3} (x_j + x_k) + \frac{2}{3} \geq \frac{4}{3} \cdot \frac{14}{15} - 36\delta - \frac{28}{15} \cdot \epsilon + 54\delta\epsilon.$$

Since $x_i + x_k \leq 1$,

$$\frac{4}{3} x_i \geq \frac{11}{45} - 36\delta - \frac{28}{15} \cdot \epsilon + 54\delta\epsilon,$$

which is impossible for $x_i \leq \epsilon$ since $\epsilon < 1/50$ and $\delta < 1/200$. 

We are now ready to prove the theorem.

**Proof of Theorem 3.** Fix the set of players $N$. Let $U$ be the uniform distribution over games in $G$. Since $N$ is fixed, we think of $U$ as a distribution over characteristic functions and write $v \sim U$.

Suppose that we draw coalitions $S_1, \ldots, S_m$ from $D$, and $v$ from $U$. Let $\mathcal{A}(\{S_1, v(S_1)\}, \ldots, \{S_m, v(S_m)\})$ be the payoff allocation returned by the given algorithm $\mathcal{A}$ on this input. Consider the event $E$ that occurs when $\mathcal{A}(\{S_1, v(S_1)\}, \ldots, \{S_m, v(S_m)\})$ is in the $(\epsilon, \delta)$-probably approximate nucleolus of the game $(N, v)$. We wish to upper-bound the probability of $E$.

To this end, instead of drawing $v$ from $U$ directly, it will be useful to use the following generative process. First, decide whether it holds that $v(S_i) = 0$ for all $i = 1, \ldots, m$; call this event $\mathcal{F}$. If $\mathcal{F}$ occurred, condition $U$ on $\mathcal{F}$ and draw $v$ from this posterior distribution. As we will see shortly, there is no need to explicitly define the process for the case where $\mathcal{F}$ did not occur.

Denoting the complement of $\mathcal{F}$ by $\tilde{\mathcal{F}}$, it holds that

$$\Pr[E] = \Pr[E \mid F] \cdot \Pr[F] + \Pr[E \mid \tilde{F}] \cdot \Pr[\tilde{F}] \leq \Pr[E \mid F] + \Pr[\tilde{F}].$$

(3)

Since for every $S_1, \ldots, S_m$, the probability of drawing $v$ from $U$ such that $\mathcal{F}$ occurs is the same by symmetry, we can compute $\Pr[F]$ by reversing the coin flips, first drawing $v$ and then $S_1, \ldots, S_m$. Only three of the $(\frac{n}{3}+1)$ coalitions of size $n/3 + 1$ have non-zero value; therefore

$$\Pr[\tilde{F}] = 1 - \left(1 - \frac{3}{(n/3+1)}\right)^m < 1/10,$$

where the inequality holds for $n \geq 9$ and $m \leq \frac{1}{6} \cdot 2^{n/3+1}$.

As for $\Pr[E \mid F]$, by Lemma 1 at least half of the games in $G$ (or, equivalently, at least half of the corresponding characteristic functions) are in the support of $U$ conditioned on $F$. But by Lemma 3, the payoff allocation $\mathcal{A}(\{S_1, v(S_1)\}, \ldots, \{S_m, v(S_m)\})$ can be in the $(\epsilon, \delta)$-probably approximate nucleolus of at most one of the $(\binom{n}{3})$ equivalence classes. It follows that

$$\Pr[E \mid F] \leq \frac{2}{(\binom{n}{3})} < 1/10.$$  

(5)

Plugging Equations (4) and (5) into Equation (3), we conclude that $\Pr[E] < 1/5$.

To recap, when drawing $S_1, \ldots, S_m$ from $D$ and $v$ from $U$, the probability that the output of $\mathcal{A}$ is in the $(\epsilon, \delta)$-probably approximate nucleolus of $G = (N, v) \in \mathcal{G}$ is at most $1/5$. But since this is true for a random game $G \in \mathcal{G}$, there must exist a game $G^* \in \mathcal{G}$ where the same is true when only drawing $S_1, \ldots, S_m$ from $D$. That is, $m$ samples are insufficient to compute a payoff allocation in the $(\epsilon, \delta)$-probably approximate nucleolus with probability at least $1 - \Delta$ for $\Delta < 4/5$. 

**4. Empirical Results**

In this section, we empirically evaluate the conclusions of our theoretical results (which are worst case in nature). We also provide comparisons between the least core and the Shapley value that shed some light on the former’s potential advantages.

Before delving into the experiments, a remark about our implementation of the least core is in order. An astute reader might have noticed that the least core is a set of solutions. To break ties, we specifically select a payoff allocation in the least core with the smallest $L^2$ norm. This is known as the egalitarian least core.
4.1. Feature Attribution

Our first set of experiments deals with feature attribution. To define a cooperative game, we take the players to be features and the value of a coalition of features to be the test accuracy of a logistic regression classifier that only has access to those features. We employ three real-world datasets of different domains from the UCI machine learning dataset directory (Dua & Graff, 2017): house (classifying the party of Congressmen based on their votes on key issues), medical (predicting the presence of breast cancer based on features of digitized images), and chemical (classifying the origin of wine based on its chemical analysis). All three datasets have 10–14 features, which makes it computationally feasible to train a logistic regression classifier on all possible subsets of features.

**Approximation Viability.** To empirically verify Theorem 1 from Section 3 (which deals with the probable least core), we sample a small fraction of coalitions uniformly at random from all possible coalitions, and compute the least core by restricting Equation (1) (with the L2 objective) to these coalitions. We then determine what fraction of all coalitions satisfy the least core constraints with respect to the true deficit $\epsilon^\star$ – that gives us $\text{accuracy} 1 - \delta$, which, in turn, leads to $\delta$-probable least core. To obtain error bars, we repeat this sampling and computation ten times. As can be seen in Figure 1, even with a small fraction of sampled coalitions, the resultant allocations are $\delta$-probable least core allocations with very small $\delta$.

Theorem 3, by contrast, asserts that many samples are needed to compute the probably approximate nucleolus. Since it is a worst-case result, though, one might wonder whether it holds in practice. To show that it does, we apply the same methodology as above. Specifically, we sample a fraction of all possible coalitions uniformly at random, and compute the nucleolus on these samples. We then determine what fraction of coalitions $(1 - \delta)$ satisfy the $(\epsilon, \delta)$-probably approximate nucleolus constraints with $\epsilon = 0.01$. As can be seen in Figure 2, in this case $\delta$ is large, even when a significant fraction of all coalitions are sampled.

**Maximum Deficit.** By definition, the maximum deficit $\epsilon^\star$ under the least core should be at most as large as that under the Shapley value. However, we wish to verify that the difference is significant in practice. To that end, we compute the least core, the Shapley value, and (as a baseline) equal payoffs on our three datasets. Figure 5 in Appendix C shows the difference between the maximum deficit of each of the solution concepts (including the least core itself) and the maximum deficit of the least core. It can be seen that there is a sizable gap between the Shapley value and the least core, considering that the maximum value of any coalition is 1. Note that no sampling (indeed, no randomness) is involved in this experiment.

**Standard Deviation.** On each of our three datasets, we compute the empirical standard deviation of payoff allocations given by the least core and the Shapley value (again no sampling is involved). Interestingly, we observe that the least core has significantly higher standard deviation and may thus be considered more discriminating; see Figure 3.

4.2. Data Valuation

Our second set of experiments deals with data valuation. We use the standard configuration of the experiment of Ghorbani & Zou (2019), in which 100 data points are sampled...
5. Discussion

Through our paper, we hope to (i) demonstrate that the least core can be approximated in a computationally tractable manner, and (ii) question the broad usage of the Shapley value with the hope of invoking further discussion on when and why one solution concept is to be preferred.

To elaborate on the second point, as we mentioned in Section 1, the core is more consistent with people’s intuition for how payoff should be divided than the Shapley value (Williams, 1988). A completely different issue is that usage of the Shapley value has almost always been justified through its four axiomatic properties (Cohen et al., 2007; Strumbelj & Kononenko, 2010; Datta et al., 2016; Lundberg & Lee, 2017; Chen et al., 2019). In this aspect, the egalitarian least core satisfies (i) efficiency (ii) symmetry and (iii) null player, only differing from the Shapley value in satisfying the stability axiom as opposed to linearity.

It is unclear why linearity should always be strictly preferred. Indeed, an $e_1$-core under coalition function $v_1$ and an $e_2$-core under coalition function $v_2$ can be combined into an allocation that satisfies the $(e_1+e_2)$-core under coalition function $v_1+v_2$ (though certainly the least core could be better than just summing the least core allocations across the two games). The necessity of linearity is commonly justified by defining a cooperative game for each test point with the coalitional value being the model accuracy with respect to that point. Hence, the overall accuracy of the test set is the sum of Shapley values. However, it is unclear why one cannot simply define one cooperative game with the coalitional value being the model accuracy with respect to the entire test set. Contrast this with the stability condition, which (when the core is nonempty) ensures non-deviation of every possible coalition of players.

Another point of difference we wish to highlight is that the (egalitarian) least core can be considered to be “harsher” than the Shapley value in terms of assigning payoffs of zero. That is, if a player is a null player under the Shapley value (does not contribute anything to any coalition), it will be assigned a payoff of 0 under the least core, but not necessarily the other way around. This is apt as the null player axiom is rarely satisfied; for instance, rarely is it the case that a feature or a data point represented by player $i$ is such that $v(\emptyset \cup \{i\}) = v(\emptyset) = 0$. Moreover, this may prove to be useful in enforcing sparsity when there is multicolinearity. Indeed, practitioners have observed that feature attribution through the Shapley value can output spuriously large influence scores (Bhatt et al., 2019).

To conclude, our theoretical and empirical results suggest that the least core is a plausible alternative means of doing credit assignment in machine learning. Currently, it appears that virtually all papers on feature attribution and data valuation use the Shapley value for that purpose. In light of the extensive applications of the core in economics and beyond, and the evidence for its potential advantages, we believe that machine learning researchers and practitioners should take a close look at this alternative approach.
References


Kopelowitz, A. Computation of the kernels of simple games and the nucleolus of N-person games. RM 31, Department of Mathematics, the Hebrew University of Jerusalem, 1967.


Appendix

A. Proof of Theorem 1

This proof is a direct extension of the proof of Theorem 1 of Balkanski et al. (2017). Like them, we employ the following known lemmas (Shalev-Shwartz & Ben-David, 2014).

Lemma 4. Let $H$ be a function class from $\mathcal{X}$ to $\{−1, 1\}$, and let $f$ be the true underlying function. If $H$ has VC-dimension $d$, then with

$$m = O\left(\frac{d + \log \left(\frac{1}{\delta}\right)}{\delta^2}\right)$$

i.i.d. samples $x^1, \ldots, x^m \sim D$,

$$\Pr_{x \sim D}\left[|h(x) - f(x)| - \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}_{h(x^i) \neq f(x^i)} \right] \leq \delta$$

for all $h \in H$ and with probability $1 - \Delta$ over the samples.

Lemma 5. The function class $\{x \mapsto \text{sign}(w \cdot x) : w \in \mathbb{R}^n\}$ has VC-dimension $n$.

We now turn to the proof. Given a coalition $S$ sampled from $D$, we convert it into a vector $y^S = (x^S, -v(S), 1)$ where $x^S_i = 1$ if $i \in S$ and $x^S_i = 0$ otherwise.

Consider a linear classifier $h$ define by $w^h = (z, 1, e)$ where $z \in \mathbb{R}^n$ and $e \in \mathbb{R}$. If $\text{sign}(w^h \cdot y^S) = 1$ then $\sum_{i \in S} z_i - V(S) + e \geq 0$. And if there exist a linear classifier $h$ that satisfies this property for all coalitions $S \in 2^N$, and in addition $z$ is efficient, then it represents a payoff allocation in the e-core. This allows us to define a class of functions that contains the e-core for all $e$. This class is:

$$\mathcal{H} = \left\{ y \mapsto \text{sign}(w \cdot y) : w = (z, 1, e), z \in \mathbb{R}^n, e \in \mathbb{R}, \sum_{i=1}^{n} z_i = v(N) \right\}.$$

This class $\mathcal{H}$ is a subset of the class of all linear classifiers of dimension $n + 2$ and thus, by Lemma 5, it has VC-dimension at most $n + 2$.

Now, suppose that we run the linear program (1) on our samples $S_1, \ldots, S_m$, which gives us a payoff allocation $\hat{z}$ and a value $\hat{e}$. Define the corresponding classifier $\hat{h}$; notice that $\hat{h}(y^S) = 1$ for all $i = 1, \ldots, m$. In addition, let $z^*$ be a payoff allocation in the least core, and $e^*$ the required subsidy, and define the corresponding classifier $f^*$. It holds that $f^*(y^S) = 1$ for all $S \in 2^N$.

By Lemma 4 we have uniform convergence for all classifiers with probability $1 - \Delta$, and in particular for $\hat{h}$ it holds that

$$\Pr_{S \sim D}\left[\sum_{i \in S} \hat{z}_i - v(S) + e^* \geq 0\right] \geq \Pr_{S \sim D}\left[\sum_{i \in S} \hat{z}_i - v(S) + \hat{e} \geq 0\right]$$

$$1 - \Pr_{S \sim D}\left[\text{sign}(w^h \cdot y^S) = -1\right]$$

$$1 - \Pr_{S \sim D}\left[\hat{h}(y^S) \neq f^*(y^S)\right]$$

$$1 - \left(\Pr_{S \sim D}\left[\hat{h}(y^S) \neq f^*(y^S)\right] - \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}_{\hat{h}(y^{S_i}) \neq f^*(y^{S_i})}\right)$$

$$\geq 1 - \delta$$

where the first transition holds because $\hat{e} \leq e^*$ and the fourth transition holds because $\hat{h}$ and $f^*$ agree on $S_1, \ldots, S_m$. $\square$

B. Proof of Theorem 2

This proof directly extends the proof of Theorem 5 of Balkanski et al. (2017). Like them, we use the following result (Shalev-Shwartz & Ben-David, 2014).
Lemma 6. Let \( \mathcal{H} = \{ w : \| w \|_1 \leq B \} \) be the hypothesis class, and \( Z = X \times Y \) be the examples domain. Suppose \( D_Z \) is a distribution over \( Z \) s.t. \( \| x \|_\infty \leq R \). Let the loss function \( \ell : \mathcal{H} \times Z \to \mathbb{R} \) be of the form \( \ell(w, (x, y)) = \phi((w, x), y) \) and \( \phi : \mathbb{R} \times Y \to \mathbb{R} \) is such that for all \( y \in Y \), the scalar function \( a \to \phi(a, y) \) is \( \rho \)-Lipschitz and such that \( \max_{a \in [-BR, BR]} \| \phi(a, y) \| \leq c \). Then for any \( \Delta \in (0, 1) \), with probability of at least \( 1 - \Delta \) over the choice of an iid sample of size \( m \) of \( (x_1, y_1), ..., (x_m, y_m) \):

\[
\mathbb{E}_{(x, y) \sim D_Z} [\ell(w, (x, y))] \leq \frac{1}{m} \sum_{i=1}^{m} \ell(w, (x^i, y^i)) + 2\rho BR \sqrt{\frac{2 \log(2d)}{m}} + c \sqrt{\frac{2 \log(2/\Delta)}{m}}.
\]

for all \( w \in \mathcal{H} \).

We also require the observation that if an \( (\epsilon, \delta) \)-probably approximate least core holds in expectation, then it is likely to hold.

Lemma 7. For any \( \epsilon > 0, \delta < 1 \) and \( \epsilon \)-core allocation \( x \) computed from samples,

\[
\mathbb{E}_{S \sim D} \left[ \frac{1 - \sum_{i \in S} z_i + \epsilon}{v(S)} \right] \leq \frac{\epsilon \delta}{1 + \epsilon} \Rightarrow \Pr_{S \sim D} \left[ \sum_{i \in S} z_i + \epsilon \geq v(S) \right] \geq 1 - \delta.
\]

Proof. Recall Markov’s inequality: for \( a > 0 \), random variable \( X \geq 0 \),

\[
\Pr[X \leq a] \geq 1 - \frac{\mathbb{E}[X]}{a}.
\]

To use it, let \( a = \frac{\epsilon \delta}{1 + \epsilon} \) and define a nonnegative random variable

\[
X = \left[ 1 - \frac{\sum_{i \in S} z_i + \epsilon}{v(S)} \right]_+.
\]

Then event \( X \leq a \) is such that

\[
X \leq a \Leftrightarrow 1 - \frac{\sum_{i \in S} z_i + \epsilon}{v(S)} \leq \frac{\epsilon}{1 + \epsilon}
\]

\[
\Leftrightarrow \sum_{i \in S} z_i + \epsilon \geq \frac{1}{1 + \epsilon} v(S)
\]

\[
\Leftrightarrow \sum_{i \in S} z_i + \epsilon + \frac{\epsilon}{1 + \epsilon} v(S) \geq v(S)
\]

\[
\Leftrightarrow \sum_{i \in S} z_i + \epsilon \geq v(S)
\]

\[
\Leftrightarrow \sum_{i \in S} z_i + \epsilon^* \geq \epsilon \geq v(S)
\]

where the penultimate step uses \( v(S) \leq 1 \) for all \( S \subseteq N \), and the last step uses that \( \epsilon^* \geq \epsilon \) since \( \epsilon \) is the least core value obtained from only a sample of all coalitional constraints.

We conclude that

\[
\Pr \left[ \sum_{i \in S} z_i + \epsilon^* + \epsilon \geq v(S) \right] \geq \Pr[X \leq a] \geq 1 - \frac{\mathbb{E}[X]}{a} \geq 1 - \frac{\delta a}{a} = 1 - \delta.
\]

\[\blacksquare\]

Turning to the theorem’s proof, in order to use Lemma 6, we begin by bounding the \( L_1 \) norm of every allocation and \( \epsilon \) in the \( \epsilon \)-core to obtain \( B \).
Therefore by Lemma 6, 

\[ ||(z, e)||_1 = v(N) + e. \] 

This holds because \( z_i \geq 0 \) for all \( i \in N \) and, by efficiency, \( ||z||_1 = v(N) \). Therefore:

\[ ||(z, e)||_1 = v(N) + e \leq v(N) + \max_S v(S) \leq 2 \max_S v(S) \]

Then, we can take our hypothesis class to be:

\[ \mathcal{H} = \{ z \in \mathbb{R}^n : ||z||_1 \leq 2 \max_S v(S) \} \]

Given \( S \sim D \), define the corresponding \( x^S = (\frac{1}{v(S)}, \frac{1}{v(S)}, \ldots) \) and the label to be \( y^S = 1 \). This allows us define to \( D_Z \) to be the uniform distribution over all \( (x^S, y^S) \) pairs. Next, suppose we obtain \( m \) samples \( S_1, \ldots, S_m \) from \( D \), the uniform distribution over all coalitions, we may again run the linear program (1) on the \( m \) samples, which gives us a payoff allocation \( \hat{z} \) and a value \( \hat{e} \). We take our classifier to be of the form \( w = (\hat{z}, \hat{e}) \) and we may define its loss \( \ell \) to be:

\[ \ell(w, (x^S, y^S)) = \ell\left((\hat{z}, \hat{e}), \left(\left(\frac{1}{v(S)}, \frac{1}{v(S)}, \ldots\right), y^S\right)\right) = \left[y^S - (\hat{z}, \hat{e}) \cdot \left(\frac{1}{v(S)}, \frac{1}{v(S)}, \ldots\right)\right]_+ = \left[1 - \sum_{i \in S} \hat{z}_i + \hat{e}\right]. \tag{6} \]

Now, we may utilize Lemma 6 with the remaining variables being \( R = \frac{1}{\min_{S \not\in \emptyset} v(S)}, B = 2 \max_S v(S), \phi(a, y) = [y - a]_+, \rho = 1 \) and \( c = 1 + 2\tau \). This is legal because, ignoring the empty set, by definition of \( x^S \), \( ||x^S||_\infty \leq \frac{1}{\min_{S \not\in \emptyset} v(S)}. \) By definition of the hypothesis class, \( ||(z, e)||_1 \leq 2 \max_S v(S) \) for all \( (z, e) \in \mathcal{H} \). \( \phi(a, y) = [y - a]_+ \) is 1-Lipschitz as:

\[
[y - a_1]_+ - [y - a_2]_+ = \max\{y - a_1, 0\} - \max\{y - a_2, 0\} \\
= \frac{|y - a_1| + y - a_1 - |y - a_2| + y - a_2}{2} \\
= \frac{|y - a_1| - |y - a_2| + a_2 - a_1}{2} \\
\leq \frac{|a_2 - a_1|}{2}
\]

Lastly, because our example domain \( Z \) is such that \( \mathcal{Y} = \{1\} \). We may obtain upper bound \( c \):

\[ c = \max_{a \in [-BR, BR]} |\phi(a, y)| = \max_{a \in [-BR, BR]} |1 - a|_+ \leq (1 - BR) = 1 + BR = 1 + 2\tau \]

Moreover, since for all \( S_i \) in our sample it holds that \( \sum_{i \in S_i} \hat{z}_i + \hat{e} \geq v(S) \), Equation (6) implies that

\[ \frac{1}{m} \sum_{i=1}^m \ell\left((\hat{z}, \hat{e}), \left(\left(\frac{1}{v(S)}, \frac{1}{v(S)}, \ldots\right), 1\right)\right) = 0. \]

Therefore by Lemma 6,

\[ E_{(x, y) \sim D}[\ell(w, (x, y))] = E_{S \sim D} \left[ \left[1 - \sum_{i \in S} \hat{z}_i + \hat{e}\right] \right]_+ \leq 0 + 2 \cdot 2 \tau \sqrt{\frac{2 \log(2(n + 1))}{m}} + (1 + 2\tau) \sqrt{\frac{2 \log(2/\Delta)}{m}} \]

Using Lemma 7, we need the number of samples \( m \) to be such that

\[ 4\tau \sqrt{\frac{2 \log(2(n + 1))}{m}} + (1 + 2\tau) \sqrt{\frac{2 \log(2/\Delta)}{m}} \leq \frac{\delta\epsilon}{1 + \epsilon}, \]

and we get that

\[ O\left(\frac{\tau^2 (\log n + \log \left(\frac{1}{\Delta}\right))}{\epsilon^2 \delta^2}\right) \]

samples suffice.
C. Additional Experimental Results

Figure 5: Relative difference between different solution concepts’ largest deficits and the least core’s largest deficit