Bounding Acceptability Degrees and Eliciting Initial Weights in Gradual Argumentation

Extended Abstract

Nir Oren
University of Aberdeen
Aberdeen, United Kingdom
n.oren@abdn.ac.uk

Bruno Yun
Universite Claude Bernard Lyon 1, CNRS, Ecole Centrale de Lyon, INSA Lyon, Université Lumière Lyon 2, LIRIS, UMR5205
Villeurbanne, France
bruno.yun@abdn.ac.uk

ABSTRACT

Many semantics for abstract weighted argumentation assume that each argument is associated with a numerical initial weight. However, eliciting these initial weights poses several challenges: (1) accurately providing a specific numerical value is often difficult, and (2) individuals frequently confuse initial weights with acceptability degrees in the presence of other arguments. To address these issues, we propose an elicitation pipeline that allows a user to specify their believed final acceptability degree intervals for each argument. We can determine which portion (if any) of these intervals are rational, refining the intervals, or restoring rationality when the intervals are irrational. This allows us to ultimately identify possible initial weights for each argument.

KEYWORDS

Argumentation; Gradual Semantics; Inverse Problems

ACM Reference Format:
Nir Oren and Bruno Yun. 2026. Bounding Acceptability Degrees and Eliciting Initial Weights in Gradual Argumentation: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/NSCC6791

1 INTRODUCTION

Building on Dung's seminal work [7], many approaches to argumentation are *abstract*, treating arguments as atomic entities which interact with each other [3]. Often, such an abstract argumentation framework is modelled as a graph, with arguments encoded in the nodes, and inter-argument interactions (such as attacks, supports [11], sets of attacking arguments [14], etc.) captured by edges.

Various enhancement to the basic model have been proposed including assigning weights to arguments [2]. Alternatives to Dung's semantics have also been considered, with *weighted gradual argumentation* [1, 5, 6] replacing the strict notion of an argument with a degree of acceptance. Therefore, arguments begin with some initial weight and — based on the structure of the graph and a semantics — a final acceptability degree for each argument is computed. For example, the weighted h-categoriser semantics computes the final acceptability degree of an argument x (denoted $\sigma^{\mathsf{Hbs}}_{\mathbf{wAF}}(x)$), with initial weight x^0, whose attackers are in $Att(x)$ as the fixed-point (i.e. $t \to \infty$) of x^t, following the formula: $x^t = x^0 / \left(1 + \sum_{y \in Att(x)} y^{t-1}\right)$.

While having advantages over classical semantics, such an approach assumes that the initial weights, and final acceptability degrees, are precise. However, humans often struggle to give precise values. Rather, humans often think in intervals. For example, the Intergovernmental panel on climate change associates the term "likely" with a confidence rating of between 66 and 90% [10, pg.3].

Furthermore, in elicitation contexts, it is often hard to obtain people's initial weights about an argument as they usually fail to disambiguate between an argument's final acceptability degree and its initial weight. For example, [12] found that individuals assign lower strength to reinstated arguments, suggesting that individuals report acceptability degree instead of initial weights when reasoning with argumentation.

Moreover, when eliciting the argument strength of multiple individual argument (which *should* be equivalent to its initial weight), one runs the risk of introducing bias, e.g., if asking someone about arguments a, b and c, the order in which they are introduced may make the person aware of these arguments which may modify the strength they ascribe to them. It is thus critical to be able to identify initial argument weights rather than final acceptability degrees.

The two issues described above motivate the current work. More specifically, we ask whether — given an interval of final acceptability degrees for arguments — is this interval rational? In other words, we ask whether the final acceptability degrees can actually be obtained from a valid assignment of initial weights to arguments. We also consider whether this interval is in some sense minimal and whether it is possible to correct any irrationality. Finally, we also provide an implementation of our system[1].

2 WORKING WITH INTERVALS

Given the problem highlighted above, we propose a framework where one can indicate an interval for acceptability degrees, from which appropriate final acceptability degrees can be drawn, and from which possible initial weights can be identified. We envision the following process: (1) an individual provides an interval in which they believe the final acceptability degree lies for every

[1]A user interface demonstrating the system (with the link to source code) is available at https://weight-elicitator.vercel.app/.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/NSCC6791

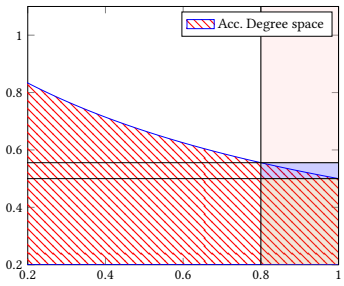

Figure 1: The x / y-axes are the degrees of a / b. The hatched area is valid acceptability degrees. Vertical lines are at $x_1 = 0.8$ and $x_2 = 1$. Horizontal lines are at $y_1 = 0.5$ and $y_2 = 5/9$.

argument. (2) If the input provided is *rational*, i.e., it is possible to find initial weights between 0 and 1 for every argument so as to obtain final acceptability degrees within the provided interval, then we *refine* the interval, narrowing the intervals down as much as possible to remove acceptability degrees which cannot be obtained from initial weights without discarding valid solutions. (3) If, on the other hand, the interval is *irrational*, i.e., one cannot find valid initial weights to yield any final acceptability degrees within the interval, then we must correct the irrationality. Doing so involves using a strategy to shift the interval bounds to make them rational.

Our approach utilises *interval constrained argumentation frameworks* (ICAFs). An ICAF is a triple of arguments $\mathcal{A}$ and attacks $\mathcal{R}$ (i.e., a standard argumentation framework) together with an interval function I such that $\forall a \in \mathcal{A}, I(a) \subseteq [0, 1]$. We illustrate these notions with the following example.

Example 1. *Consider an ICAF $(\mathcal{A}, \mathcal{R}, I)$, where $\mathcal{A} = \{a, b\}$ and $\mathcal{R} = \{(a, b)\}$, and let $U = 5/9$ and $L = 0.5$. In Figure 1, we represent the acceptability degree space of a and b (below the blue curve) w.r.t. the weighted h-categoriser. For every point (d_a, d_b) in this area, there exists a weighting function w (and corresponding* $\mathbf{wAF} = (\mathcal{A}, \mathcal{R}, w)$ *s.t. $\sigma^{\mathsf{Hbs}}_{\mathbf{wAF}}(a) = d_a$ and $\sigma^{\mathsf{Hbs}}_{\mathbf{wAF}}(b) = d_b$. Assume that $I(a) = [0.8, 1.0]$.*

- *If $I(b) \in [x, L]$, for $0 \leq x \leq L$, then the ICAF is fully rational as for every (d_a, d_b), $d_a \in I(a)$, and $d_b \in I(b)$, there exists a weighting function w s.t. $\sigma^{\mathsf{Hbs}}_{\mathbf{wAF}}(a) = d_a$ and $\sigma^{\mathsf{Hbs}}_{\mathbf{wAF}}(b) = d_b$. Any such point (d_a, d_b) will lie within the brown rectangle.*
- *If $I(b) = [x, y]$, for $L < y, 0 \leq x \leq U$ and $x \leq y$, then the ICAF is rational, as the point $(0.8, x) \in I(a) \times I(b)$ is in the brown or blue area but it is not fully rational as the point $(1, y) \in I(a) \times I(b)$ is not in the acceptability degree space (but still lies inside the blue area).*
- *Finally, if $I(b) = [x, y]$ s.t. $U < x \leq y$, then the ICAF is irrational because every point in $I(a) \times I(b)$ lies in the pink area (and thus above the blue curve).*

Note that intervals on acceptability degree cannot be directly converted into intervals for initial weights. This is due to the interdependencies between arguments in the final acceptability degree computation process. For example, if $I(a) = [0.8, 1]$ and $I(b) = [0, 0.5]$, one would think that the initial weights of a lies between 0.8 and 1 whereas the one of b would lie within 0 and 1. However, if $w(a) = 0.8$ and $w(b) = 1$, we would have $\sigma^{\mathsf{Hbs}}_{\mathbf{wAF}}(b) \notin I(b)$.

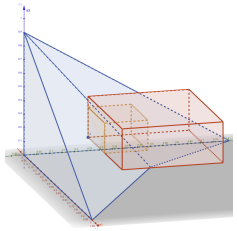

Figure 2: ε-refinement.

If we have full rationality we can identify initial weights between 0 and 1 for any final acceptability degree within an interval. However, with rationality alone, there is some combination of final acceptability degrees within the intervals which will be achievable. We may wish to measure how close to fully rational an ICAF is under some semantics. This notion is captured via ϵ-rationality.

Intuitively, ϵ measures how much the *upper* end of the interval needs to be reduced to ensure that the entire interval is rational. We can then *refine* the interval to include only appropriate solutions, to yield a so-called ϵ-refinement. This is illustrated in Figure 2. In this illustration, the blue shaded area represents the space of final acceptability degrees which originate from valid initial weights, while the large cuboid represents the intervals of all arguments (each argument's final acceptability degree is along one of the figure's axes). The original interval (larger cuboid) is then maximally shrunk, or refined by shifting the corner furthest away from the origin in such a way so as to ensure that the resultant set of intervals (the ϵ-refinement), shown by the smaller cuboid in the figure.

3 RELATED WORK

In [8], the authors introduce credal support argumentation frameworks where each argument is associated with a credal set and an imprecise base score is computed. While they only consider a support relation, they show how to compute the imprecise strength of arguments (as an interval) and study theoretical properties. The epistemic approach to probabilistic argumentation [9] aims to determine valid probabilities for arguments given some properties, similar to our intervals. In the context of fuzzy argumentation, legal argument weights can be computed according to the approach of [13]. However, the properties of the interval are not explicitly considered, nor are correcting systems which do not comply with the underlying properties. Enforcement in abstract argumentation (e.g., [4]) considers how argumentation frameworks can be modified to guarantee an argument's status and refinement can be viewed similarly in a weighted gradual semantics context.

4 CONCLUSIONS

We outlined how intervals of final acceptability degrees can be refined with respect to gradual argumentation and initial weights. Applications include knowledge engineering and assisting humans in reasoning. Extending the work described here, we have also investigated heuristics to correct irrationality. unlike refinement, these algorithms move the point *closest* to the origin to perform such a correction. Due to space constraints we refer the reader to https://arxiv.org/abs/2502.07452v1 for further details.

REFERENCES

[1] Leila Amgoud and Jonathan Ben-Naim. 2013. Ranking-Based Semantics for Argumentation Frameworks.. In *Scalable Uncertainty Management - 7th International Conference, SUM 2013, Washington, DC, USA, September 16-18, 2013. Proceedings*. 134–147. https://doi.org/10.1007/978-3-642-40381-1_11

[2] Leila Amgoud and Jonathan Ben-Naim. 2018. Weighted Bipolar Argumentation Graphs: Axioms and Semantics.

[3] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. 2011. An introduction to argumentation semantics. *Knowledge Eng. Review* 26, 4 (2011), 365–410. https://doi.org/10.1017/S0269888911000166

[4] Ringo Baumann, Sylvie Doutre, Jean-Guy Mailly, and Johannes Peter Wallner. 2021. Enforcement in Formal Argumentation. *IfColog Journal of Logics and their Applications (FLAP)* 8, 6 (July 2021), 1623–1678. https://hal.science/hal-03541704 ISBN : 978-1-84890-371-5.

[5] Philippe Besnard and Anthony Hunter. 2001. A logic-based theory of deductive arguments. *Artif. Intell.* 128, 1-2 (2001), 203–235. https://doi.org/10.1016/S0004-3702(01)00071-6

[6] Elise Bonzon, Jérôme Delobelle, Sébastien Konieczny, and Nicolas Maudet. 2018. Combining Extension-Based Semantics and Ranking-Based Semantics for Abstract Argumentation.. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR 2018, Tempe, Arizona, 30 October - 2 November 2018*. 118–127. https://aaai.org/ocs/index.php/KR/KR18/paper/view/18067

[7] Phan Minh Dung. 1995. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. *Artif. Intell.* 77, 2 (1995), 321–358. https://doi.org/10.1016/0004-3702(94)00041-X

[8] Mariela Morveli Espinoza, Juan Carlos Nieves, and Cesar A. Tacla. 2023. A Gradual Semantics with Imprecise Probabilities for Support Argumentation Frameworks. In *Proceedings of the 21st International Workshop on Non-Monotonic Reasoning co-located with the 20th International Conference on Principles of Knowledge Representation and Reasoning (KR 2023) and co-located with the 36th International Workshop on Description Logics (DL 2023), Rhodes, Greece, September 2-4, 2023 (CEUR Workshop Proceedings, Vol. 3464)*, Kai Sauerwald and Matthias Thimm (Eds.). CEUR-WS.org, 84–93. https://ceur-ws.org/Vol-3464/paper9.pdf

[9] Anthony Hunter, Sylwia Polberg, Nikolas Potyka, Tjitze Rienstra, and Mathhias Thimm. 2021. *Handbook of Formal Argumentation*. Vol. 2. College Publications, Chapter Probabilistic argumentation: A Survey.

[10] IPCC. 2023. Climate Change 2023: Synthesis Report. , 35–115 pages. https://doi.org/10.59327/IPCC/AR6-9789291691647

[11] Antonio Rago, Francesca Toni, Marco Aurisicchio, and Pietro Baroni. 2016. Discontinuity-Free Decision Support with Quantitative Argumentation Debates.. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016*. 63–73. http://www.aaai.org/ocs/index.php/KR/KR16/paper/view/12874

[12] Iyad Rahwan, Mohammed Iqbal Madakkatel, Jean-François Bonnefon, Ruqiyabi Naz Awan, and Sherief Abdallah. 2010. Behavioral Experiments for Assessing the Abstract Argumentation Semantics of Reinstatement. *Cogn. Sci.* 34, 8 (2010), 1483–1502. https://doi.org/10.1111/J.1551-6709.2010.01123.X

[13] Jiachao Wu, Hengfei Li, Nir Oren, and Timothy J. Norman. 2016. Gödel Fuzzy Argumentation Frameworks. In *COMMA (Frontiers in Artificial Intelligence and Applications, Vol. 287)*. IOS Press, 447–458.

[14] Bruno Yun, Srdjan Vesic, and Madalina Croitoru. 2020. Ranking-Based Semantics for Sets of Attacking Arguments. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 3033–3040. https://doi.org/10.1609/AAAI.V34I03.5697