Towards Weak Assumption-Based Argumentation

Lydia Blümel

Artificial Intelligence Group, University of Hagen, Universitätsstraße 11, 58097 Hagen

Abstract

Assumption-Based Argumentation (ABA) is a versatile non-monotonic reasoning formalism that derives arguments from a set of defeasible assumptions via some given inference rules. Like many other formal reasoning tools, ABA also allows the derivation of contradictions, giving rise to direct and indirect self-contradicting arguments. For Abstract Argumentation, a closely related formalism, semantics based on weak admissibility have been shown to solve this problem while otherwise behaving nicely. Instead of the classic defense by counter-attack, they use a recursive notion of admissibility to check the validity of each incoming attack. We explore how this approach can be realized in the more expressive ABA formalism. Our proposal is to use Abstract Bipolar SETAFs as an in-between to take advantage of the abstract intuition behind weak admissibility without a loss in expressiveness on the side of ABA.

Keywords

Argumentation, Knowledge Representation Languages

1. Introduction

In recent years, Computational Argumentation has become one of the leading approaches for Explainable Artificial Intelligence (XAI), while still being rooted firmly in Symbolic AI and Reasoning. Debating different possible theories and outcomes is an integral part of human intelligence gathering and decision making, so when it comes to deriving, communicating, verifying and justifying artificial knowledge, argumentation is a natural method to use. In particular, structured argumentation formalisms like Assumption-Based Argumentation (ABA) [1] are applied in key areas like planning [2], health care [3] and causal discovery [4]. One advantage of ABA is its high expressiveness despite its relatively simple structure. An Assumption-based Argumentation Framework (ABAF) consists of a set of defeasible assumptions, a set of rules for making inferences from these assumptions in a given formal language and a contrary function that tells us what constitutes a counter-argument against each assumption. In contrast to Abstract Argumentation [5], where only atomic arguments and the attacks between them are considered, ABA allows us to model set attacks and, in the case of Non-Flat ABA, which also permits the derivation of assumptions from other assumptions, even set support between assumptions. However, the attack relation is not explicit in ABA but results from applying the rules exhaustively to generate arguments from assumptions and attacking the assumptions used for each argument according to the contrary function. Due to this, contradictions and inconsistencies within the rules can be hard to pinpoint when evaluating an ABAF and indeed, existing ABA semantics fall short when it comes to the treatment of, e. g. , self-contradicting assumptions. A promising solution for this problem in Abstract Argumentation are semantics based on weak admissibility [6]. They rely on a recursive defense notion to distinguish reasonable from unreasonable attackers. Our work aims to introduce weak admissibility to the significantly more expressive ABA formalism in a meaningful way, taking into account the many facets of self-contradictions in ABA. So far, we managed to define weak admissibility for Flat [7] as well as Non-Flat ABA [8] indirectly via a suitable abstract representation. In this proposal, we give a quick overview on our progress to date, starting with an example to motivate the study of weak semantics for ABA in general. We then give the definition of weak admissibility for Abstract Bipolar SETAFs (BSAF) and show how the combination of a BSAF-representation with weak admissibility maintains enough expressiveness to treat paradoxical rules in ABA satisfactory. We conclude with ideas for future work.

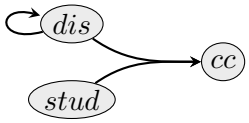
Doctoral Consortium of the 22nd International Conference on Principles of Knowledge Representation and Reasoning (KR 2025 DC), November 11-17, 2025, Melbourne, Australia

lydia.bluemel@fernuni-hagen.de (L. Blümel)

2. Motivating Problem

In both Structured and Abstract Argumentation, a key concept is the notion of *admissibility*. Loosely speaking, admissibility formalizes that, in a given debate, an acceptable set S of arguments should (i) not contain internal conflicts and (ii) be able to refute arguments raised against S. However, as pointed out in several works [5, 9, 10], this requirement is often too strong in real-world argumentation scenarios, especially in the presence of paradoxical arguments as in the following example.

Example 2.1. *Suppose our agent participates in a debate about climate change. The first argument brought forward is that "Climate change is happening due to mankind emitting carbon dioxide." (yielding an assumption for climate change, cc). Another argument confirms this, stating that "According to numerous studies, climate change is happening." (yielding an assumption in favor of studies, $stud$). However, another participant counters this by arguing that "I read on social media that everything written on the internet is false." (yielding an assumption in favor of distrusting information, dis). If we distrust every information on the internet, this together with the fact that the studies on climate change are available online constitutes a collective attack from dis and $stud$ towards cc: if both of these assumptions are accepted, we have to disregard cc. On the other hand, dis is a self-attacking argument, because if everything on the internet is false, then also this same information found on social media. We obtain the following attack structure.*



Clearly, in this scenario we would like to disregard the self-attacking argument that represents distrusting reasonable information. However, no commonly agreed semantics for ABAFs can handle this in a satisfactory way.

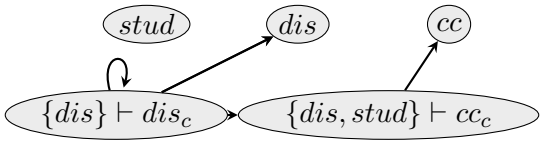
As we said in our introduction, ABA is a suitable formalism for modeling a variety of argumentation scenarios like this example due to its simple structure but more than sufficient expressiveness. Before we go into further detail, we therefore give a short formal introduction to ABA. Based on a deductive system $(\mathcal{L}, \mathcal{R})$, where $\mathcal{L}$ is a set of sentences and $\mathcal{R}$ a set of inference rules, an ABA framework (ABAF) is a tuple $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$, where $\mathcal{A}$ is a set of assumptions from which we infer according to the rules $\mathcal{R}$, and $c : \mathcal{A} \to \mathcal{L}$ a contrary function we use to derive attacks on assumptions. In order to conduct defeasible reasoning with an ABAF $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$ we consider *arguments* that can be built by applying rules to assumptions. More precisely, we consider the fact, that some $p \in \mathcal{L}$ can be derived from a set of assumptions $A \subseteq \mathcal{A}$ according to the rules $\mathcal{R}$ an *(ABA) argument*, and denote this by $A \vdash p$. Furthermore, if p is the contrary $c(b)$ of some assumption $b \in \mathcal{A}$ we say the set A *attacks* any set $B \subseteq \mathcal{A}$ with $b \in B$.A set of assumptions A is conflict-free if it does not attack itself; admissible if it is conflict-free and defends itself, i. e. attacks all of its attackers. Consider the introductory example.

Example 2.2. *We can model our introductory Example 2.1 with the ABAF $\mathcal{A} = \{stud, dis, cc\}$, $\mathcal{L} = \mathcal{A} \cup \{a_c \mid a \in \mathcal{A}\}$,*
contraries $c(a) = a_c$ for each assumption and rules
$\mathcal{R} = \{(dis_c \leftarrow dis), (cc_c \leftarrow dis, stud)\}$.

The assumption cc is attacked by the set $\{dis, stud\}$ and cannot be defended, because the only attacker of that set is the assumption dis itself.

At first, we tried to use weak admissibility for AFs in ABA out of the box: the idea was to simply evaluate the argumentation graph F_D corresponding to an ABAF D. Given an ABAF $D = (\mathcal{L}, \mathcal{R}, \mathcal{A}, c)$, the corresponding argumentation graph is the tuple $F_D = (A_D, R_D)$ where A_D is the set of all ABA arguments $S \vdash p$ with $p \in \mathcal{L}$ derived from a set of assumptions $S \subseteq \mathcal{A}$ according to the rules $\mathcal{R}$ and attacks $(S \vdash p, T \vdash q)$ if p is the contrary of some $t \in T$. Now if an argument attacks itself in an abstract argumentation framework, it can be ignored under weak admissibility and is neither accepted nor relevant for the acceptance of other arguments. This direct approach, however, does not work as we demonstrate next.

Example 2.3. *Instantiating the ABAF from Example 2.2 as an AF yields the following argumentation graph :*



Instead of disregarding the self-attacking assumption dis*, weak admissibility only allows us to disregard the* argument *"*$c(dis) \leftarrow dis$*" which explicates that this assumption is self-attacking. Consequently,* dis *and* $stud$ *are accepted whereas* cc *is not. This is surprisingly far away from what we want to achieve; after all, weak admissibility handles (abstract) self-attackers quite well.*

3. Progress to date

In [7] we propose to use SETAFs [11] as an alternative representation for ABAFs which have the necessary expressiveness to properly identify self-attacks and odd cycles in an ABAF. A SETAF is a tuple $SF = (A, R)$, where A is a set of arguments and R is a set of *collective* attacks (T, h) from a set of arguments $T \subseteq A$ to a single argument $h \in A$. Using the reduct notion for SETAFs from [12] we define weak admissibility on SETAFs by generalizing the definition for AFs in a natural way. By this, we indirectly define weak admissibility for ABA. An ABAF is represented by a SETAF where the assumptions are the set of arguments and an attack from a set of assumptions T to some assumption h is simply a collective attack between the respective arguments corresponding to these assumptions. Now a set of assumptions is weakly admissible, if the corresponding set of arguments in the SETAF-representation is weakly admissible. We can indeed successfully capture the motivating example with this representation. However, SETAFs are only sufficient as a representation for *Flat* ABAFs [13], i. e. ABAFs where assumptions cannot be derived by the rules. To overcome this limitation, we further generalized the notion of weak admissibility to Bipolar SETAFs in this years paper [8]. Bipolar SETAFs [14] combine the ideas underlying argumentation frameworks with collective attacks (SETAFs) [11] and bipolar argumentation frameworks (BAFs) [15, 16, 17]. Instead of only considering an attack relation, there is also a notion of support. BSAFs can model *collective* attacks and supports.

Definition 3.1. *A bipolar set-argumentation framework (BSAF) is a tuple* $F = (A, R, S)$*, where* A *is a finite set of arguments,* $R \subseteq 2^A \times A$ *is the attack relation and* $S \subseteq 2^A \times A$ *is the support relation.*

We restrict ourselves to finite BSAFs, those that have finitely many arguments.

Definition 3.2. *Given a BSAF* $F = (A, R, S)$ *and a set* $E \subseteq A$ *of arguments. With*

$$supp_F(E) := E \cup \{h \in A \mid \exists\, (T, h) \in S : T \subseteq E\}$$

we call $cl_F(E) = \bigcup_{i \geq 1} supp_F^i(E)$ *the* closure *of* E*;* E *is* closed *if* $cl_F(E) = E$*.*

We denote by E_R^+ the set of all arguments attacked by E and define $E_R^\oplus = E \cup E_R^+$. A set $E \subseteq A$ is conflict-free ($E \in cf(F)$) if it does not attack itself; *defends* $a \in A$ if for each closed set $E' \subseteq A$ attacking a, E attacks E'; E *defends* E' if E defends each $a \in E'$.

A conflict-free set E is admissible ($E \in adm(F)$) iff E is closed and defends itself.

Due to space limitations we omit the introduction of the standard semantics. Before we can give the definition of weakly admissible semantics for this framework, we first need a suitable notion of reduct.

Definition 3.3. *Given a BSAF* $F = (A, R, S)$ *and* $E \subseteq A$*, the* E-reduct of F *is the BSAF* $F^E = (A^E, R^E, S^E)$*, with*

$$A^E = A \setminus (cl(E)_R^\oplus)$$

$$
\begin{aligned}
R^E = \{(T \setminus cl(E), t) \mid \exists h \in cl(E)^+_R : (T,h) \in S,\\
t \in T \cap A^E\} \cup\\
\{(T \setminus cl(E), h) \mid T \cap cl(E)^+_R = \emptyset,\\
(T,h) \in R^E, h \in A^E\}\\
S^E = \{(T \setminus cl(E), h) \mid T \cap cl(E)^+_R = \emptyset,\\
(T,h) \in S, h \in A^E\}
\end{aligned}
$$

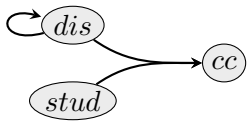
Note that some design choices had to be made for handling collective attacks and supports. Intuitively, an attack/support is removed in the reduct, whenever one attacking/supporting assumption is attacked by the set E in question, otherwise it is restricted to the assumptions remaining in the reduct. Furthermore, we add self-attacks in the reduct for sets of assumptions that support an argument attacked by E. The E-reduct for BSAFs gives us the tools to generalize weak admissibility. Note that the definition is recursive, but well-defined as in each recursion step the reduct contains fewer arguments and we deal only with finite BSAFs.

Definition 3.4. *Let $F = (A, R, S)$ be a BSAF, $E \subseteq A$ a set of arguments, and $F^E = (A^E, R^E, S^E)$ its E-reduct. Then E is called* weakly admissible *in F ($E \in adm^w(F)$) iff*

1. *$E \in cf(F)$, E closed and*
2. *for each $(T,h) \in R$ with $h \in E$, and $T \cap E^+_R = \emptyset$ it holds $\nexists E' \in adm^w(F^E)$ s.t. $T \cap A^E \subseteq E'$.*

When applying this definition to our introductory example, we can now ignore the joint attack involving the self-attacker and accept the argument that climate change is happening.

Example 3.5. *Recall our introductory Example 2.1 about climate change. The instantiated BSAF BF_D is given as follows.*



We have $adm^w(BF_D) = \{\emptyset, \{stud\}, \{cc\}, \{stud, cc\}\}$.

So our motivating example is indeed handled as desired now. To demonstrate that weak admissibility behaves well in general, let us consider the paradoxical rule principle from [8]. As the name indicates, the principle involves attacks and supports in the ABAF that we deem paradoxical. We introduce the concept (in terms of BSAFs as they resemble ABA attacks and supports) below.

Definition 3.6. *Given a BSAF $F = (A, R, S)$. An attack $r = (T,h) \in R$ is* paradoxical *iff $T \neq \emptyset$ and for every $t \in T$ there is a $T' \subseteq T, T' \neq \emptyset$ s.t.*

- *there exists $(T', t) \in R$ and $h \notin T'$.*

A support $s = (T,h) \in S$ is paradoxical *iff $T \neq \emptyset$ and for every $t \in T$ there is a $T' \subseteq T, T' \neq \emptyset$ s.t.*

- *there exists $(T', t) \in R$.*

We want weak admissibility to be stable under removal of paradoxical rules, thus the principle states the following requirement.

(**PRS**) Paradoxical Attacks/Supports: Removing a paradoxical attack r or support s does not alter the models of F, i. e. $adm^w(F) = adm^w(F')$ where $F' = (A, R \setminus \{r\}, S)$ (resp. $F' = (A, R, S \setminus \{s\})$).

Proposition 3.7. *The weakly admissible semantics for ABA satisfies the Principle of Paradoxical Attacks/Supports.*

With this result from [8], we have shown that weak admissibility under the BSAF-instantiation captures self-conflicting sets of assumptions in general ABA in a natural way.

4. Future Work

Successfully formulating weak admissibility for general ABA provides valuable insights towards a native reduct notion for ABA and an ABA-semantics satisfying long-standing rationality postulates like non-interference [18]. Another natural next step would be to introduce weak admissibility to other structured argumentation formalism, e. g. ASPIC [19]. Since reasoning with weak admissibility comes with a high computational complexity [20] the practical feasibility of using weak admissibility for reasoning with ABA should be checked. In [21] we investigate a class of semantics that approximate the weakly preferred semantics in the abstract setting. As these semantics are also based on the notion of reduct they could be utilized to make weak admissibility realistic for applications which use ABA as a reasoning tool [2, 3].

Acknowledgments

The research reported here was partially supported by the Deutsche Forschungsgemeinschaft (grant 550735820).

Declaration on Generative AI

The author has not employed any Generative AI tools.

References

[1] A. Bondarenko, P. M. Dung, R. A. Kowalski, F. Toni, An abstract, argumentation-theoretic approach to default reasoning, Artif. Intell. 93 (1997) 63–101. URL: http://dx.doi.org/10.1016/S0004-3702(97)00015-5. doi:10.1016/S0004-3702(97)00015-5.

[2] X. Fan, A temporal planning example with assumption-based argumentation, in: Proc. PRIMA 2018, 2018, pp. 362–370.

[3] K. Cyras, T. Oliveira, A. Karamlou, F. Toni, Assumption-based argumentation with preferences and goals for patient-centric reasoning with interacting clinical guidelines, Argument Comput. 12 (2021) 149–189. URL: https://doi.org/10.3233/AAC-200523. doi:10.3233/AAC-200523.

[4] F. Russo, A. Rapberger, F. Toni, Argumentative causal discovery, in: P. Marquis, M. Ortiz, M. Pagnucco (Eds.), Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024, 2024. URL: https://doi.org/10.24963/kr.2024/88. doi:10.24963/KR.2024/88.

[5] P. M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games, Artif. Intell. 77 (1995) 321–358. doi:10.1016/0004-3702(94)00041-X.

[6] R. Baumann, G. Brewka, M. Ulbricht, Revisiting the foundations of abstract argumentation - semantics based on weak admissibility and weak defense, in: Proc. AAAI 2020, AAAI Press, 2020, pp. 2742–2749.

[7] L. Blümel, M. König, M. Ulbricht, Weak admissibility for ABA via abstract set-attacks, in: P. Marquis, M. Ortiz, M. Pagnucco (Eds.), Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024, 2024. URL: https://doi.org/10.24963/kr.2024/17. doi:10.24963/KR.2024/17.

[8] M. Berthold, L. Blümel, A. Rapberger, On strong and weak admissibility in non-flat assumpion-based argumentation, in: Proceedings of the 22nd International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Melbourne, Australia. November 11-17, 2025, 2025. URL: tobepublished.

[9] P. Dondio, L. Longo, Beyond reasonable doubt: A proposal for undecidedness blocking in abstract argumentation, Intelligenza Artificiale 13 (2019) 123–135.

[10] R. Baumann, G. Brewka, M. Ulbricht, Shedding new light on the foundations of abstract argumentation: Modularization and weak admissibility, Artif. Intell. 310 (2022) 103742. URL: https://doi.org/10.1016/j.artint.2022.103742. doi:10.1016/J.ARTINT.2022.103742.

[11] S. H. Nielsen, S. Parsons, A generalization of Dung's abstract framework for argumentation: Arguing with sets of attacking arguments, in: Proc. ArgMAS 2006, volume 4766 of *Lecture Notes in Computer Science*, Springer, 2006, pp. 54–73.

[12] W. Dvořák, M. König, M. Ulbricht, S. Woltran, Principles and their computational consequences for argumentation frameworks with collective attacks, J. Artif. Intell. Res. 79 (2024) 69–136. URL: https://doi.org/10.1613/jair.1.14879. doi:10.1613/JAIR.1.14879.

[13] M. König, A. Rapberger, M. Ulbricht, Just a matter of perspective, in: Proc. COMMA 2022, volume 353 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2022, pp. 212–223. URL: https://doi.org/10.3233/FAIA220154. doi:10.3233/FAIA220154.

[14] M. Berthold, A. Rapberger, M. Ulbricht, Capturing non-flat assumption-based argumentation with bipolar setafs, in: P. Marquis, M. Ortiz, M. Pagnucco (Eds.), Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, KR 2024, Hanoi, Vietnam. November 2-8, 2024, 2024. URL: https://doi.org/10.24963/kr.2024/12. doi:10.24963/KR.2024/12.

[15] C. Cayrol, M. Lagasquie-Schiex, On the acceptability of arguments in bipolar argumentation frameworks, in: L. Godo (Ed.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 8th European Conference, ECSQARU 2005, Proceedings, volume 3571 of *Lecture Notes in Computer Science*, Springer, 2005, pp. 378–389. URL: https://doi.org/10.1007/11518655_33. doi:10.1007/11518655_33.

[16] L. Amgoud, C. Cayrol, M.-C. Lagasquie, P. Livet, On bipolarity in argumentation frameworks, International Journal of Intelligent Systems 23 (2008) 1–32.

[17] M. Ulbricht, N. Potyka, A. Rapberger, F. Toni, Non-flat ABA is an instance of bipolar argumentation, in: M. J. Wooldridge, J. G. Dy, S. Natarajan (Eds.), Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada, AAAI Press, 2024, pp. 10723–10731. URL: https://doi.org/10.1609/aaai.v38i9.28944. doi:10.1609/AAAI.V38I9.28944.

[18] A. Borg, C. Straßer, Relevance in structured argumentation, in: J. Lang (Ed.), Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, ijcai.org, 2018, pp. 1753–1759. URL: https://doi.org/10.24963/ijcai.2018/242. doi:10.24963/IJCAI.2018/242.

[19] S. Modgil, H. Prakken, The aspic+ framework for structured argumentation: a tutorial, Argument & Computation 5 (2014) 31–62.

[20] W. Dvorák, M. Ulbricht, S. Woltran, Recursion in abstract argumentation is hard - on the complexity of semantics based on weak admissibility, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, AAAI Press, 2021, pp. 6288–6295. URL: https://doi.org/10.1609/aaai.v35i7.16781. doi:10.1609/AAAI.V35I7.16781.

[21] L. Blümel, M. Thimm, Approximating weakly preferred semantics in abstract argumentation through vacuous reduct semantics, in: P. Marquis, T. C. Son, G. Kern-Isberner (Eds.), Proc. KR 2023, 2023, pp. 107–116.