Assessing a Formal Model of Reflective Equilibrium

Technical Report (v1.0)

Andreas Freivogel Sebastian Cacean

2024-08-11

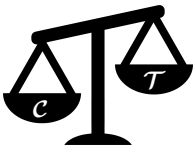
Table of contents

1 Preface

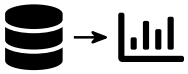
1.1 Abstract

In philosophy, and especially in ethics, reflective equilibrium (RE) is often considered a powerful method for obtaining beliefs that mutually support each other, are justified by evidence, and are backed by good reasons. Beisbart, Betz, and Brun (2021) have introduced a formal model of reflective equilibrium based on the theory of dialectical structures Betz (2013), which they use as a methodological tool to understand the method of reflective equilibrium better. This report is an outcome of the research project 'How far does Reflective Equilibrium Take us? Investigating the Power of a Philosophical Method' and summarizes the findings of assessing the model thoroughly by numerical investigation. We simulate RE processes for a broad spectrum of model parameters and initial conditions and use four different model variants (including the original model). We analyze the dependence of simulation results on different parameters and assess the models' conduciveness towards consistency, and ability to reach global optima and full RE states. The results show that the models' behaviour depends crucially on the specifics of the simulation setup (e.g., the sentence pool size and α weights). We can, therefore, not draw any general conclusions about the overall performance of the model variants. Rather, the specifics of the context in which an RE model is used must be considered to choose a specific model. Finally, we identify some critical knowledge gaps we cannot close with this report that call for further research into RE modelling.

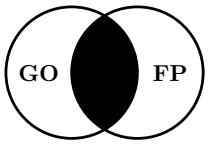
1.2 Content



In Chapter 2, we introduce the formal model of reflective equilibrium of Beisbart, Betz, and Brun (2021) together with three variations of the original model that have been included in this report. We motivate the metrics for model validations that guide our assessment. Finally, we describe the ensemble of RE simulations that has been generated by the computer implementation of the formal model of RE.



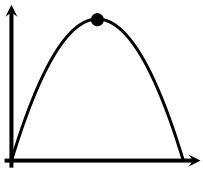
In Chapter 3, we present general results about the ensemble of RE simulation that form the basis of this report. They help to understand the model better, and they ease the interpretation of salient results, subsequently.



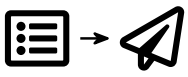
In Chapter 4, we provide results concerning the overlap of two outputs produced by the model: global optima and fixed points. They represent the static aspect of equilibrium states and the dynamic aspect of equilibration processes in RE, respectively.

In Chapter 5, we present results concerning the attainment of full RE states which meet the highest standards for RE outputs. Full RE states represent outputs that can be understood to be justified by RE.

In Chapter 6, we analyse different aspects of consistency pertaining to the outputs of the formal model. Commonly, consistency is considered to be a necessary requirement for coherence.

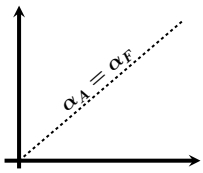


In Chapter 7, we display outcomes regarding maximal values of measures of RE desiderata that guide the selection of states. This part of the analysis aims to foster understanding about the trade-offs in the formal model of RE.



In Chapter 8, we summarize the main outcomes of the report and provide an outlook to promising lines of future research.

Appendices



In Appendix A , we prove analytic results about linear model variants. These results explain the salient behaviour of linear model variants that occurs throughout the report.

$$\{s\} = \{s\}$$

In Appendix B, we analyse data with respect to the attainment of "trivial" outcomes, i.e. states that consist of a single commitment paired with a singleton theory.

In Appendix C, we discuss alternative systematicity measures by analytical means. In view of shortcomings of the original systematicity measure, we evaluate the newly proposed measures in view of various desiderata for such measures.

1.3 Reproducibility

All findings and the underlying data can be reproduced by using the Python implementation of the model. The data that the model produced can be found here. For each chapter you will find here a Jupyter notebook whose execution produces all analysis results. For more specific instructions of how to reproduce all findings, please refer to the github repo of this report.

1.4 Licence

Figure 1.1: CC BY 4.0

1.5 Citation

BibTex citation:

```
@article{freivogel_assessing_2024,
  title = {Assessing a {{Formal Model}} of {{Reflective Equilibrium}}},
  author = {Freivogel, Andreas and Cacean, Sebastian},
  year = {2024},
  month = august,
  doi = {10.5281/zenodo.13294165},
  langid = {english},
  url = {https://re-models.github.io/re-technical-report/},
}
```

For attribution, please cite this works as, for instance:

Freivogel, A., & Cacean, S. (2024). *Assessing a Formal Model of Reflective Equilibrium.* https://doi.org/10.5281/zenodo.13294165

1.6 Credits

This report is part of the research project 'How far does Reflective Equilibrium Take us? Investigating the Power of a Philosophical Method' (SNSF grant 182854 and German Research Foundation grant 412679086). Earlier versions of it were discussed on several occasions with all members of the project. We thank, in particular, Claus Beisbart, Gregor Betz, Georg Brun, Alexander Koch and Richard Lohse for their helpful comments, which helped to improve this report considerably. Finally, the authors acknowledge support by the state of Baden-Württemberg through the joint high-performance computer system bwHPC.

2 Introduction

Beisbart, Betz, and Brun (2021) have introduced a formal model of reflective equilibrium based on the theory of dialectical structures (Betz 2010, 2013), which they use as a methodological tool to better understand the method of reflective equilibrium and to assess its potential to yield justified epistemic states. Their discussion of the model is mainly based on an illustrative example. An assessment of how the model behaves under a broader spectrum of circumstances went beyond the scope of their work.

This report summarizes findings of assessing the RE model more thoroughly by numerical investigation. We simulated RE processes for a broad spectrum of model parameters, initial conditions and with different model variants. We compare simulation outcomes of three model variants to the ones of the original model and analyze the dependence of simulation results on different parameters.[1]

2.1 Modelling Reflective Equilibration

Reflective equilibrium is commonly understood as a method of justification, in which an epistemic subject iteratively adjusts their epistemic state in a process of equilibration until a state of reflective equilibrium is reached. In this final state, the agent's belief system is supposed to be justified to the extent that it satifies various pragmatic-epistemic objectives, e.g., (internal) coherence.

Beisbart, Betz, and Brun (2021) model this process of reflective equilibration and the underlying axiology of equilibrium states in the following way.[2]

The agent's epistemic state is modelled as a tuple $(\mathcal{C}, \mathcal{T})$, which comprises their accepted commitments $\mathcal{C}$ and a theory $\mathcal{T}$. Both are represented by sets of sentences from a finite pool of sentences $\mathcal{S}$, which is closed under negation.

The equilibration process is modelled as a mutual adjustment of the theory and the agent's commitments to improve the epistemic state as measured by an achievement function Z (see Figure Figure 2.1). The agent starts with a set of initial commitments $\mathcal{C}_0$. Then, a theory $\mathcal{T}_0$

[1] The results of Beisbart, Betz, and Brun (2021) are based on a Mathematica implementation of the model (see https://github.com/debatelab/remoma). Here, we rely on a reimplementation in Python (rethon), which can reproduce the results of the original implementation (see this notebook).

[2] For a thorough and complete description of the formal RE model, see Beisbart, Betz, and Brun (2021). The present section is based on condensed material from Freivogel (2023).

is chosen that systematizes $\mathcal{C}_0$. This initial state $(\mathcal{C}_0, \mathcal{T}_0)$ is then adjusted by searching for a new set of commitments that performs better in terms of the overall achievement Z. This process of adjusting the current epistemic state by choosing a new theory (or new commitments, respectively) goes on until no further improvement is gained any more.

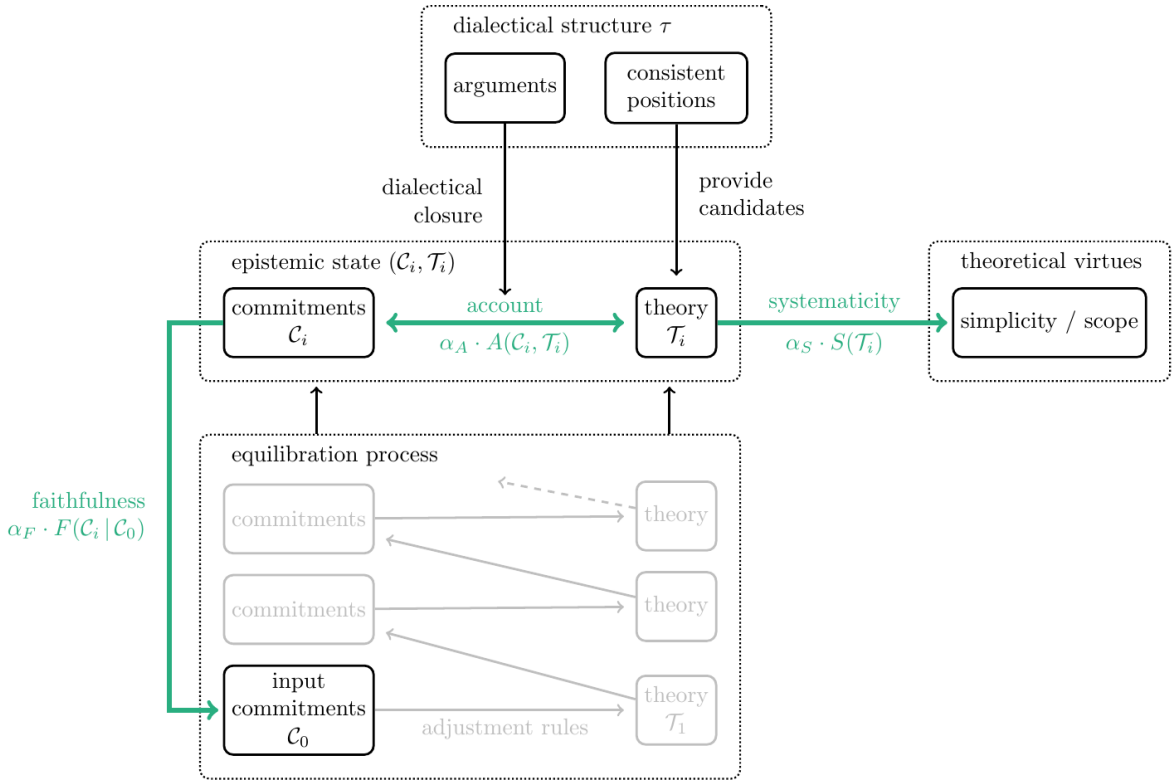


Figure 2.1: Illustrative diagram of the formal model. The epistemic state, which consists of a set of commitments and a theory, is subject to operationalized desiderata (F, S and A) for RE states (bold arrows). Rules for alternating adjustments of commitments and theory specify a process of equilibration that sets out from initial commitments. Taken from Freivogel (2023) (CC BY).

The achievement function Z models the underlying axiology and is based on the three different desiderata *faithfulness* (F), *systematicity* (S) and *account* (A). Their role is illustrated by bold arrows in Figure Figure 2.1.[3]

The desideratum of *faithfulness* demands that current commitments should not deviate too much from the initial commitments $\mathcal{C}_0$. There are two motivations for this constraint (Beisbart, Betz, and Brun 2021, 447). A resemblance of the current commitments to $\mathcal{C}_0$ contributes to the justification of the resulting state to the extent that initial commitments have some independent credibility. Additionally, the sentences in $\mathcal{C}_0$ represent a specification of the topic

[3]For formal details of all measures, see (Beisbart, Betz, and Brun 2021, 464–66).

under consideration. Deviating too much from $\mathcal{C}_0$ courts the danger of changing the topic. Faithfulness $F(\mathcal{C} \,|\, \mathcal{C}_0)$ is operationalized in the model by measuring the distance of the current commitments to the initial commitments.[4]

The role of the theory $\mathcal{T}$ is to systematize the commitments $\mathcal{C}$. Beisbart, Betz, and Brun (2021) suggest to explicate this idea by asking whether the theory implies the commitments. The account $A(\mathcal{C}, \mathcal{T})$ measures how well the commitments $\mathcal{C}$ fit to what is implied by the theory $\mathcal{T}$. More specifically, $A(\mathcal{C}, \mathcal{T})$ is based on measuring the distance between $\mathcal{C}$ and the set of $\mathcal{T}$'s implications.

To that end, we need to know how the sentences in $\mathcal{S}$ are inferentially connected. The inferential relationships are modelled by dialectical structures based on the theory of dialectical structures (Betz 2010, 2013). A dialectical structure τ is a set of deductively valid arguments $\mathcal{A}$ and their "inferential" relationships to each other. For instance, an argument with two premises s_i, s_j ($\in \mathcal{S}$) and a conclusion s_k represents the inferential relationship of s_k being implied by the conjunction of s_i and s_j.[5] Each process of reflective equilibration takes place on the background of one dialectical structure that stays fixed during the process.

The final desideratum demands that a theory does not only perform well in systematizing the commitments $\mathcal{C}$ but is generally able to systematize sentences in $\mathcal{S}$ (independent of whether they belong to the agent's epistemic state). Systematicity $S(\mathcal{T})$ measures this general inferential potential by considering the amount of $\mathcal{T}$'s implications in relation to the size of the sentence pool $\mathcal{S}$.

All three desiderata can "pull" in different directions. The resolution of such trade-offs is modelled by using a convex combination of the three measures as a one-dimensional combined measure Z for the overall epistemic quality of the agent's epistemic state:

$$Z(\mathcal{C}, \mathcal{T} \,|\, \mathcal{C}_0) = \alpha_A \cdot A(\mathcal{C}, \mathcal{T}) + \alpha_S \cdot S(\mathcal{T}) + \alpha_F \cdot F(\mathcal{C} \,|\, \mathcal{C}_0),$$

The weights α_A, α_S and α_F are real-valued numbers between 0 and 1 that sum up to 1. Different suggestions for balancing the desiderata are represented by choosing different α-weights in the achievement function Z.

The achievement function assigns to every epistemic state $(\mathcal{C}, \mathcal{T})$ an epistemic value. Epistemic states that maximize this value are called *global optima*. The evaluation of epistemic states is relative to what we can call an *epistemic situation* of an agent, i.e., a dialectical structure τ, a set of initial commitments $\mathcal{C}_0$, and a configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$. The epistemic situation captures the subject matter of inquiry, its background, and decisions to handle trade-offs between epistemic desiderata.

[4] The used distance is a weighted Hamming distance. For details, see Beisbart, Betz, and Brun (2021), 465.

[5] The arguments of a dialectical structure τ need not be formally valid, but can include "arguments that are valid *given the relevant background theories*" (Beisbart, Betz, and Brun 2021, 460). Additionally, τ does not need to codify all inferential relationships between sentences in $\mathcal{S}$ and can, in this way, model some form of bounded rationality.

2.2 Model Variations

In this report, we compare the performance of four model variants that result from a combination of two independent alterations of the original model from Beisbart, Betz, and Brun (2021) (see Table 2.1). First, we will vary the general shape of the functions A, S and F. In the original model, these functions have a quadratic form, which will be contrasted with a linear form. Second, we will compare the semi-global optimization during equilibrations steps, which is used in Beisbart, Betz, and Brun (2021), with a locally optimizing model variant.

	Quadratic shape	Linear shape
Global optimization	`QuadraticGlobalRE` (in short, `QGRE`)	`LinearGlobalRE` (in short, `LGRE`)
Local optimization	`QuadraticLocalRE` (in short, `QLRE`)	`LinearLocalRE` (in short, `LLRE`)

Table 2.1: Model variations

2.2.1 Quadratic and Linear Measures

In Beisbart, Betz, and Brun (2021), the functions A, F and S have the following shape:

$$G(x) = 1 - x^2$$

However, the quadratic term x^2 is not motivated. The linear models `LGRE` and `LLRE` will be based on $G(x) = 1 - x$ to examine the repercussions of such a variation.

2.2.2 Semi-globally and Locally Optimizing Equilibration Processes

The mutual adjustment of commitments and theories involves two types of revisions. The agent will revise their current commitments $\mathcal{C}_i$ and their current theory $\mathcal{T}_i$ in an alternating fashion. More specifically, when adjusting their commitments, the agent will search for new commitments $\mathcal{C}_{i+1}$ such that the resulting state $(\mathcal{C}_{i+1}, \mathcal{T}_i)$ performs better w.r.t. Z. Similarly, when adjusting their theory, the agent will search for a theory $\mathcal{T}_{i+1}$ such that $Z(\mathcal{C}_i, \mathcal{T}_{i+1} \,|\, \mathcal{C}_0) > Z(\mathcal{C}_i, \mathcal{T}_i \,|\, \mathcal{C}_0)$.

The equilibration process in Beisbart, Betz, and Brun (2021) is a semi-global optimization in the following way: When searching for new commitments $\mathcal{C}_{i+1}$ that improve Z, the agent can choose any set of commitments. Similarly, when searching for a new theory $\mathcal{T}_{i+1}$, the agents can choose any theory. This search strategy is computationally costly as the search space grows exponentially with the size of the sentence pool. For the same reason, it is also an unrealistic assumption about real epistemic subjects.

To solve this problem and incorporate some form of bounded rationality into the model, we can constrain the search space for adopting new commitments and theories. Instead of considering all commitments and theories, a *locally* optimizing equilibration process confines the search space to a neighbourhood of the current state.

The definition of this neighbourhood is based on an edit distance, which measures the number of changes needed to transform one set of sentences into another. Suppose the sentence pool $\mathcal{S}$ comprises three sentences and their negations—that is, $\mathcal{S} = \{s_1, s_3, s_3, \neg s_1, \neg s_2, \neg s_3\}$. Let us now consider two different sets of commitments: $\mathcal{C}_1 = \{s_1, \neg s_2\}$ and $\mathcal{C}_2 = \{s_1, s_2, s_3\}$. Suppose further that an agent adopts $\mathcal{C}_1$ as their commitments. In other words, they accept s_1, refuse s_2 and are indifferent towards s_3. Consequently, a set of commitments can be specified by describing the doxastic attitude (acceptance, refusal and indifference) towards each sentence of half the sentence pool (s_1, s_2 and s_3 in our example). The edit distance we use is defined by asking how many changes of doxastic attitudes are needed to transform one set of commitments into another. Consequently, the edit distance between $\mathcal{C}_1$ and $\mathcal{C}_2$ is 2 since we would have to change the attitude for s_2 from refusal to acceptance and for s_3 from indifference to acceptance.

We can now define the neighbourhood of depth d (in short, the d-neighbourhood) of a set of sentences S_i as the set of all sentence sets that have at most an edit distance of d to S_i.[6]

The *local* model variants `QLRE` and `LLRE` restrict the commitments and theory candidates during adjustment steps to a neighbourhood of depth $d = 1$.

To illustrate the difference between global, semi-global and local optimization, think of epistemic states $(\mathcal{C}, \mathcal{T})$ as cells on an appropriately sized, possibly non-square, chess board.[7] The unbounded, globally optimizing agent can overview the entire board at once (Figure 2.2), while a semi-globally optimizing agent can evaluate only a single row or column per adjustment step (Figure 2.3). Finally, only candidates from a small neighbourhood of the current position are available to the locally optimizing agent only during an adjustment step (Figure Figure 2.4).

2.3 Metrics for Model Validations

At the outset, a plethora of metrics could be used to examine the performance of the formal model. Let us motivate a small selection of desiderata for model validation, which we will use in the following chapters.

[6] For a sentence pool size of $2n$, the number of positions in the neighbourhood of a position is $\sum_{k=0}^{d} \binom{n}{k} \cdot 2^k$, where d denotes the neighbourhood depth. For $d = 1$, the number of positions in the neighbourhood grows linearly with the number of sentences. More specifically, for $d = 1$, the size of the neighbourhood is $2n + 1$.

[7] Note that the two-dimensional representation of the epistemic states in the subsequent figures is purely illustrative. There is no inherent linear order among positions, which can be understood as points in an n-dimensional discrete space. Similarly, the indices in the figures are not supposed to correspond to the order of commitments and theories in the evolution of the epistemic state.

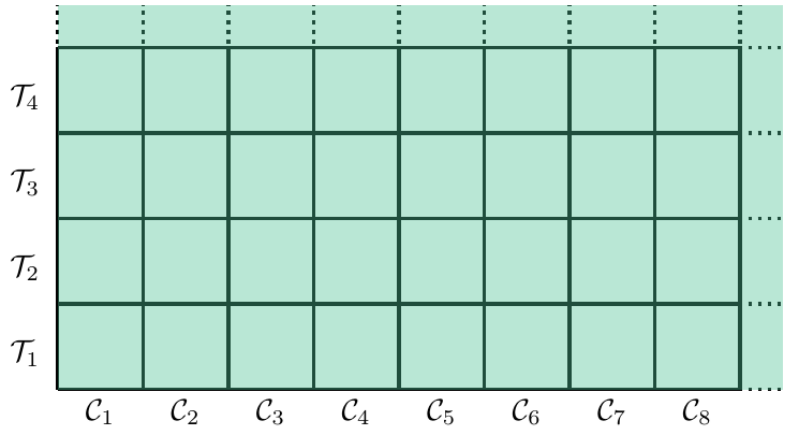


Figure 2.2: Global optimization: All epistemic states are available.

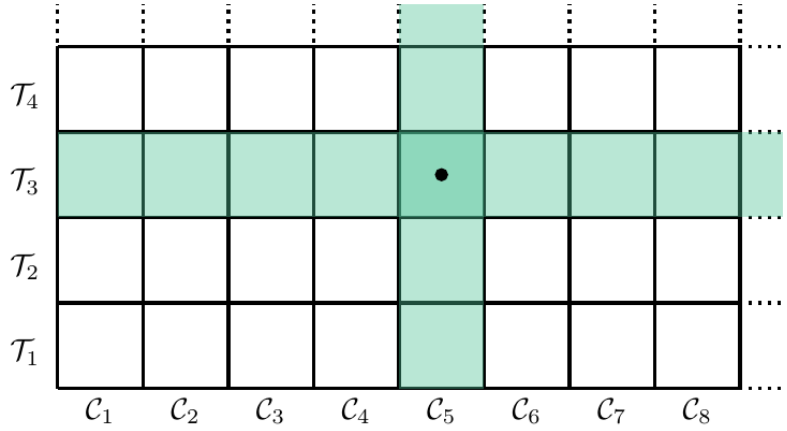


Figure 2.3: Semi-global optimization: All sets of commitments and all theories are available in an alternating fashion while the other component is held fixed.

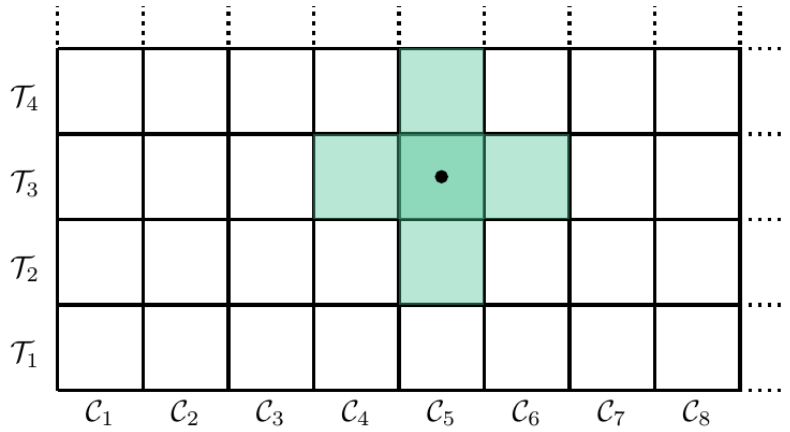


Figure 2.4: Local optimization (alternating): Available commitments(row)/theories (column) are restricted to a neighbourhood of the current state in an alternating fashion while the other component is held fixed.

The process of reflective equilibration reaches an endpoint, a so-called fixed point, if the agent arrives at an epistemic state that cannot be further improved (in terms of the achievement function) by revising their commitments or their theory, respectively (Beisbart, Betz, and Brun 2021, 450). However, such a fixed point is not necessarily a global optimum. In other words, other epistemic states might perform better w.r.t. Z.

This possible divergence of fixed points and global optima applies to locally optimizing models (`LLRE` and `QLRE`) and the semi-globally optimizing models (`LGRE` and `QGRE`). The former can get stuck in local optima since they are confined to a restricted search area for the improvement of epistemic states. However, the semi-globally optimizing models can also get stuck in local optima since they do not adjust their commitments and theories simultaneously but alternately. Consequently, we must distinguish between the axiology of the RE (as defined by the achievement function) as a static aspect of RE and the equilibration process as the procedural aspect of RE.[8]

Accordingly, several questions concerning the relationship between fixed points and global optima are relevant to the performance assessment of the model variants. In Chapter 4, we investigate whether fixed points are global optima and, conversely, whether global optima are reachable by equilibration processes.

The reached achievement of fixed points and global optima is not the only evaluative perspective on epistemic states. In other words, there are other aspects of evaluating reflective equilibria besides the desiderata of account, systemticity and faithfulness (Beisbart, Betz, and Brun 2021, 448–49).

[8]The fact that the model allows distinguishing static and dynamic aspects makes the model a fruitful foil to discuss the broader epistemological questions surrounding the method of reflective equilibrium (Beisbart, Betz, and Brun 2021, 457–58).

The most ambitious requirement demands that a theory accounts fully and exclusively for the commitments of an epistemic state. Global optima and globally optimal fixed points that additionally satisfy this criterion are called full RE states. In Chapter 5, we investigate whether and under which circumstances fixed points and global optima are full RE states. We will also analyze whether theories of global optima fully and exclusively account for their commitments.

Weaker requirements demand that fixed points or, at least, fixed point commitments are dialectically consistent—that is, consistent with respect to all inferential relationships encoded in the given dialectical structure τ. Consistency is commonly seen as a necessary condition of coherence. Achieving consistency is, therefore, of utmost importance for equilibration processes. In Chapter 6, we will assess the consistency conduciveness of the different model variants.

Finally, we will investigate whether global optima and fixed points yield extreme values in the normalized measures A, F and S (Chapter 7). The achievement function Z aggregates these measures by using weights to model trade-offs between the desiderate (e.g., give up on faithfulness to increase account). Investigating under which circumstances extreme values are achieved in A, F, and S might improve our understanding of the involved trade-offs and of the consequences of choosing specific weights.

2.4 Ensemble Description

> **Note**
>
> The results presented in this section can be reproduced with the following notebook: https://github.com/re-models/re-technical-report/blob/main/notebooks/chapter_general-props.ipynb.

The results of RE processes and their global optima depend on the following inputs:

- the model variant,
- the dialectical structure τ and the sentence pool,
- the α-weights for the achievement function and
- the set of initial commitments

Let us call a specification of these inputs a *simulation setup*.

Due to the exponential growth of candidate commitments and theories, which all have to be considered for global optima and semi-global adjustment steps in RE processes, the ensemble includes only four sentence pools with a small number of sentences (12, 14 , 16, 18). We generated 50 dialectical structures and 20 sets of random initial commitments for each sentence pool. We used every resulting configuration of dialectical structures and initial commitments

(out of $4\,000 = 4 \cdot 50 \cdot 20$ configurations) to run RE processes for every of the described model variants and for 36 α-weight configurations.

For each of the resulting 576 000 simulation setups, we calculated global optima. Note also that one simulation setup does not necessarily determine a fixed point uniquely. For every step of adjusting a theory (or a set of commitments), the subsequent theory (or set of commitments) is underdetermined if there is more than one candidate that maximizes the achievement function. In such cases, the model will randomly choose the next theory (or set of commitments) (Beisbart, Betz, and Brun 2021, 466). The Python implementation of the model allows us to track each of the resulting branches, which we did for this report.

However, for some simulation setups (2 765), we do not have simulation results. Due to reasons of computational feasibility, we had to set a cut-off point for the number of branches per simulation setup and the number of adjustment steps. Model runs that exceeded these thresholds were interrupted.[9] We chose to limit the number of branches to 400 and set the maximum number of adjustment steps to 100. Given these restrictions, the resulting ensemble comprises 4 136 547 branches.

Setups	Model variants	Sentence pool sizes	Dialectical structures	Initial commitments	α-weights (resolution)
576 000	4	12, 14, 16, 18	$4 * 50$	$4 * 20$	36 (0.1)

Table 2.2: Ensemble properties

Let us now describe the simulation setups more thoroughly.

2.4.1 α-Weights

Each α-weight was varied between 0.1 and 0.8 in steps of 0.1 (i.e., values from $0.1, 0.2, \ldots, 0.8$). Since α-weights have to satisfy $\alpha_A + \alpha_S + \alpha_F = 1$, there are 36 possible combinations of the described α-weights.

We excluded extreme values such as $\alpha_F = 0$ or $\alpha_A = 1$ since they "break" the model and lead to undesirable behaviour. For example, $\alpha_F = 0$ results in global optima that comprise all and only singleton theories and their closures as commitments.

[9] The model runs will always converge within a finite number of steps into a fixed point (see Beisbart, Betz, and Brun 2021, 467). The same could be shown for the number of branches. However, the number of branches and the length of processes can still be computationally challenging.

2.4.2 Initial Commitments

We generated a simple random sample of 20 sets of minimally consistent initial commitments for every sentence pool. While we allow initial commitments to be dialectically inconsistent—that is, inconsistent w.r.t. the inferential relationships codified in τ— they must be minimally consistent. In other words, they should not include flat contradictions (e.g., $\{s_1, s_2, \neg s_1\}$).

Let $2^{\mathcal{S}}$ be the set of all sets of minimally consistent sets of sentences from $\mathcal{S}$.[10] If $2n$ is the size of the sentence pool, then there are 3^n minimally consistent sets of sentences ($|2^{\mathcal{S}}| = 3^n$). For the generation of the used random sample, every set of commitments has the same probability of being drawn. Note that this does not translate into a uniform distribution of the number of initial commitments since the amount of sets with a specific size varies in $2^{\mathcal{S}}$. In Figure 2.5, you find the actual distribution of the initial commitments' sizes for the different sentence pools.

Roughly, 55% of the random initial commitments are dialectically consistent. This value varies slightly depending on the sentence pool (see Figure 2.6).

2.4.3 Dialectical Structures

We generated 50 random dialectical structures for each sentence pool, which codify all inferential relationships on which an RE process is based. A dialectical structure comprises arguments with an internal premise-conclusion structure and dialectical relationships between arguments. Arguments can attack or support each other (see Figure 2.7 for an example).[11]

Inferential relationships are represented in a dialectical structure τ in the following way: If the sentences $P = \{s_{i_1}, s_{i_2}, \ldots, s_{i_m}\}$ are premises of an argument in τ and s_j is its conclusion, s_j is (known to be) implied by P.

The support and attack relation are defined as follows: If an argument A supports another argument B, the conclusion of A is (known to be) equivalent to a premise of B; if an argument A attacks another argument B, the conclusion of A is (known to be) inconsistent with a premise of B.

The *inferential density* of a dialectical structure τ "can be understood as measure of the inferential constraints encoded in τ" (Betz 2013, 44) and is defined as

$$D(\tau) = \frac{n - lg(\sigma)}{n}$$

where $2n$ is the size sentence pool and σ the number of complete and consistent positions in τ

The τ-generating algorithm we used receives the following parameters as constraints:

[10] $2^{\mathcal{S}}$ is a subset of the powerset of $\mathcal{S}$, which is usually denoted by $2^{\mathcal{S}}$.

[11] The illustrated dialectical structure is one from the actual data set (with the name `tau-alpha-020`). Argument maps of all used dialectical structures can be found here.

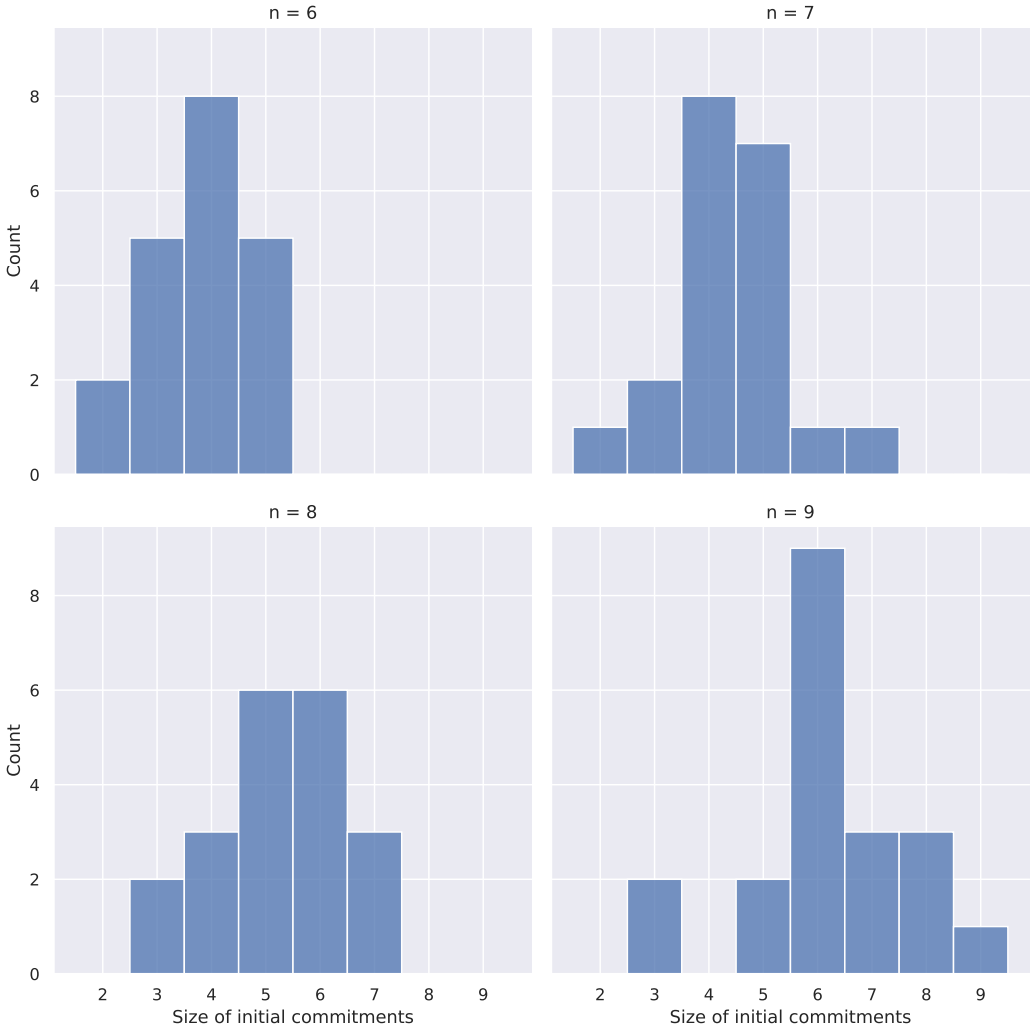


Figure 2.5: Distribution of initial commitments' sizes for different n.

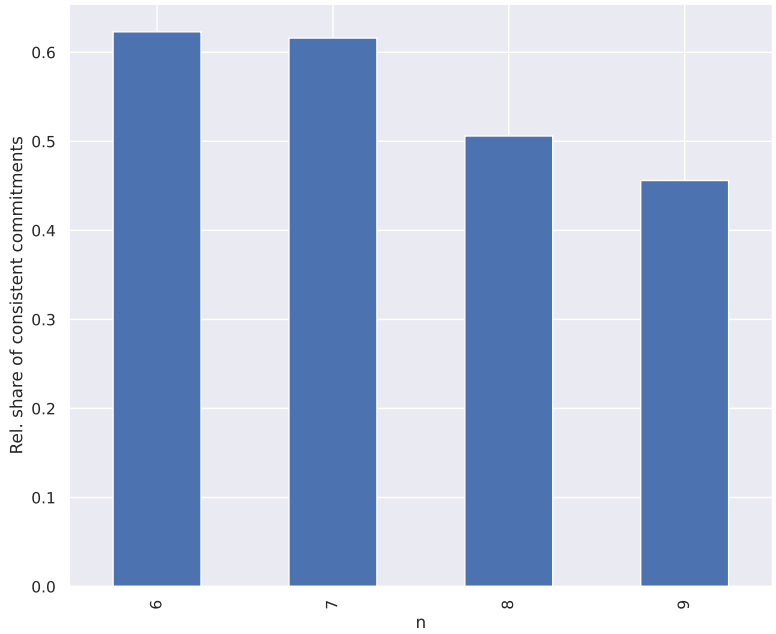


Figure 2.6: Relative share of dialectically consistent initial commitments for different n.

- the size of the sentence pool ($n \in \{6, 7, 8, 9\}$),
- an interval for the permissible number of arguments ($I_{|\tau|} = [n - 2, n + 2]$),
- the maximum number of premises per argument ($P_{n_{max}} = 2$),
- probability weights for the number of premises for arguments (i.e., weights for each $1, \ldots, P_{n_{max}}$) and
- an interval for the permissible inferential density ($I_D = [0.15, 0.5]$)

The algorithm will generate a dialectical structure by randomly constructing arguments so that the number of arguments and the inferential density fall in the specified intervals $I_{|\tau|}$ and I_D. Both properties correlate inversely: Roughly, the more arguments τ has, the higher its inferential density.

Besides the specified parameters, the algorithm will satisfy the following constraints:

- The dialectical structure is satisfiable (i.e., there is at least one dialectically consistent position on τ).
- Every sentence will be used. In other words, for every sentence $s \in \mathcal{S}$, there is an argument in τ such that s or $\neg s$ is either a premise or the conclusion of the argument.

- Arguments are not question-begging (i.e., an argument's conclusion is not in its premise set).
- Arguments are not attack-reflexive (i.e., the negation of an argument's conclusion is not in its premise set).

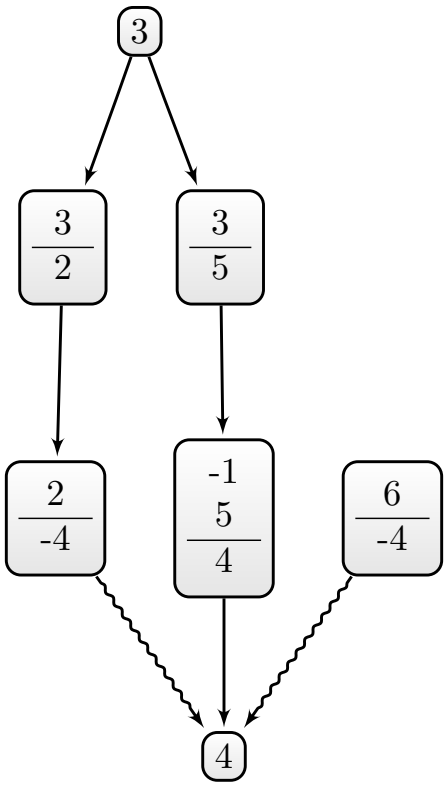


Figure 2.7: Example of a dialectical structure. Attack relations are indicated by waved-shaped arrows, and support relations by straight arrows. Numbers represent sentences from $\mathcal{S}$, and the minus sign denotes the negation of a sentence.

Figure 2.8 plots the actual distribution of inferential densities in the generated data set of all 200 dialectical structures. It shows that the inferential density is not uniformly distributed. Instead, we observe a bias towards dialectical structures with an inferential density on the lower side of I_D. This is an artefact of the τ-generating algorithm, which tries to generate an arbitrary dialectical structure satisfying the described constraints. Since it is "easier" to produce a dialectical structure with a lower inferential density, the algorithm produces dialectical structures with comparably lower values from I_D.[12]

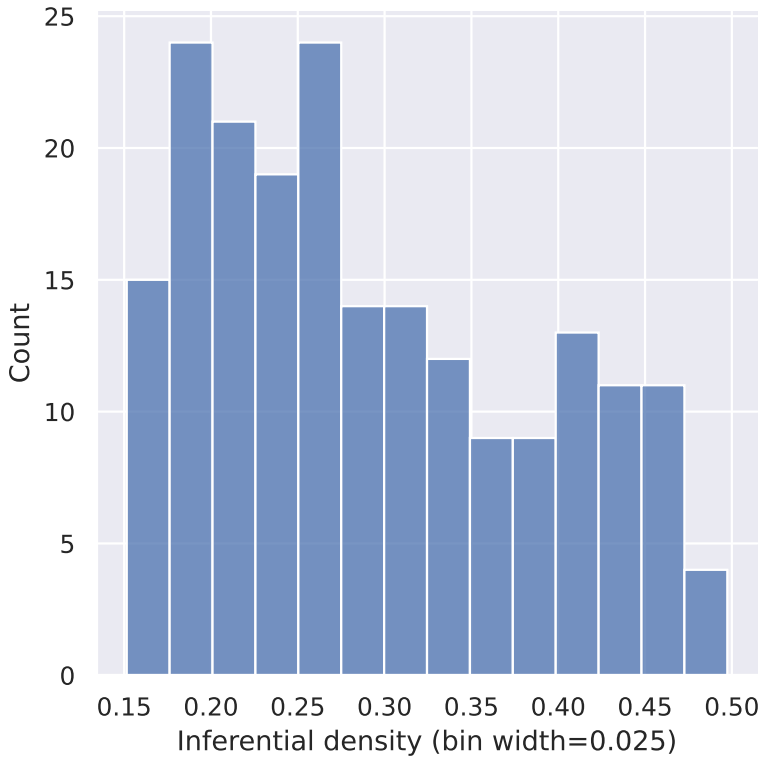


Figure 2.8: Distribution of inferential densities in the used τ-data set.

All generated dialectical structures have arguments with 1-2 premises. For each sentence pool, we used five sets of weights for the number of premises such that there are 10 dialectical structure with an expected number of premises $E(|P_\tau|)$ of 1, 10 with $E(|P_\tau|) = 1.25$, 10 with $E(|P_\tau|) = 1.5$, 10 with $E(|P_\tau|) = 1.75$ and 10 with $E(|P_\tau|) = 2$. The resulting actual distribution of the mean number of premises per argument can be seen in Figure 2.9. The increased amount of τs with only 1 and 2-premise arguments results from ceiling effects since all arguments have at least one and at most two premises.

[12]For specifics, consider the implementation.

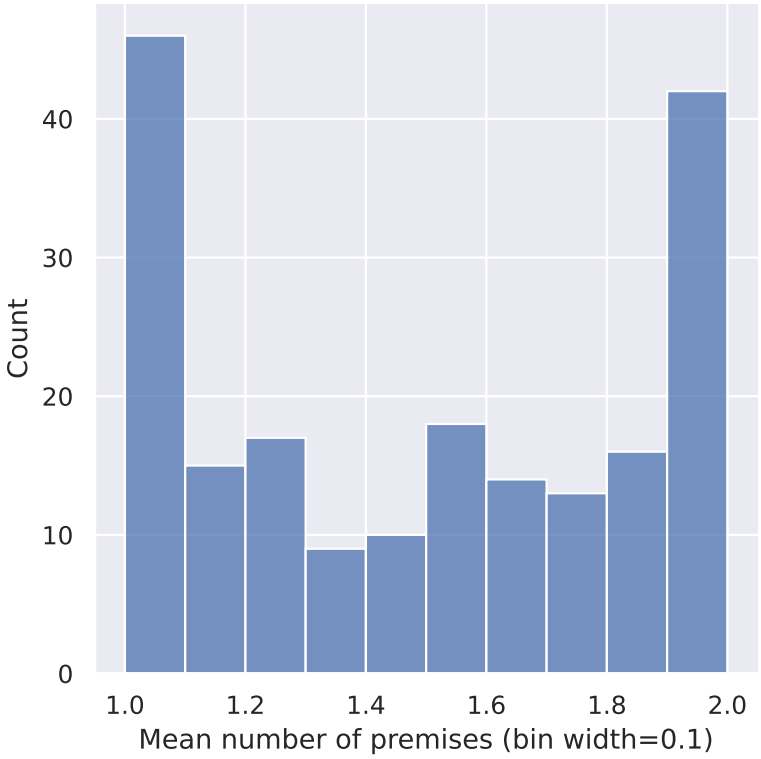


Figure 2.9: Distribution of the mean number of premises in the used τ-data set.

3 General Ensemble Properties

> **Note**
>
> The results of this chapter can be reproduced with the following Jupyter notebook: https://github.com/re-models/re-technical-report/blob/main/notebooks/chapter_general-props.ipynb.

Before analyzing how the model variants perform with regard to the described performance criteria, we will analyze some basic features of model runs that help us understand the model better. The resulting insights will (hopefully) help to understand and interpret some of the results, we will present in subsequent chapters.

In particular, we assess the overall length of processes, the (mean) step length of commitments adjustments steps, properties of global optima and the extent of branching.

Here and in the following chapters, we will assess different properties of model runs and their dependence (mainly) on the chosen model variant, the selection of α weights and the size of the sentence pool. Admittedly, other dimensions as, for instance, properties of the dialectical structures (such as inferential density) are also interesting as independent variables to assess the performance of the different models. However, we had to confine the analysis to some extent and regard the chosen dimensions as particularly important.

Since we want to compare the performance of different model variants, we have, of course, to vary the model. The variation of α-weights is important since the modeler has to choose a particular set of α-weights in a specific context. It is therefore not enough to know how the different models compare to each other on average (with respect to α-weights) but important to compare them within different confined spectra of α-weight configurations. The dependence on the size of the sentence pool is motivated by the practical restrictions to use semi-globally optimizing model variants. Due to computational complexity the use of semi-globally optizing models is feasible for small sentence pools only. However, these small sentence pools are too small to model reflective equilibration of actual real-world debates.[1] Accordingly, we are confined to use locally optimizing models in these cases. It is, therefore, of particular interest whether the observations of this ensemble study can be generalised to larger sentence pools.

[1] Note/link about Andreas' modelling of Tanja's reconstruction.

3.1 Process Length and Step Length

In the following, we understand *process length* (l_p) as the number of theories and commitment sets in the evolution e of the epistemic state, including the initial and final state.

$$\mathcal{C}_0 \to \mathcal{T}_0 \to \mathcal{C}_1 \to \mathcal{T}_1 \to \cdots \to \mathcal{T}_{final} \to \mathcal{C}_{final}$$

In other words, if $(\mathcal{T}_0, \mathcal{C}_0)$ is the initial state and $(\mathcal{T}_m, \mathcal{C}_m)$ the fixed-point state, $l_p(e) = 2(m+1)$. An equilibration process reaches a fixed point if the newly chosen theory and commitments set are identical to the previous epistemic state—that is, if $(\mathcal{T}_{i+1}, \mathcal{C}_{i+1}) = (\mathcal{T}_i, \mathcal{C}_i)$ (Beisbart, Betz, and Brun 2021, 466). Therefore, the minimal length of a process is 4. In such a case, the achievement of initial commitments and the first chosen theory cannot be further improved. Accordingly, the initial commitments are also the final commitments.

Figure 3.1 shows the distribution of process lengths, and Figure 3.2 shows the mean process length (and its standard deviation) for the different model variants dependent on the size of the sentence pool ($2n$) over all branches.

Note that Figure 3.1 counts branches of a particular length for each model. One simulation setup can result in different branches if the adjustment of commitments or theories is underdetermined. Additionally, the number of branches for a specific simulation setup can vary between different models. Consequently, the overall number of branches per model can differ. This, in turn, explains why the sum of bars varies between the subfigures of Figure 3.1 (see Section 3.3 for details).

The first interesting observation is that the semi-globally optimizing models (`QuadraticGlobalRE` and `LinearGlobalRE`) reach their fixed points quickly. Often, they adjust their commitments only once ($l_p(e) = 6$); the linear model variant (`LinearGlobalRE`) will sometimes not even adjust the initial commitments of processes ($l_p(e) = 4$). In contrast, the locally optimizing models (`QuardraticLocalRE` and `LinearLocalRE`) need significantly more adjustment steps. This difference is expected if we assume that local and global optima commitments are not often in the 1-neighbourhood of initial commitments (see Figure 3.4 and Figure 3.9). Under this assumption, the locally searching models will need more than one adjustment step to reach a global or local optimum.

Additionally, the models `QuardraticLocalRE` and `LinearLocalRE` have a much larger variance in process lengths than the models `QuadraticGlobalRE` and `LinearGlobalRE`.

A third observation concerns the difference in process lengths between semi-globally and locally optimizing models in terms of their dependence on the sentence pool. Figure 3.2 suggests that the process length of locally optimizing models increases with the size of the sentence pool. The semi-globally optimizing models lack such a dependence on the sentence pool size.

A possible explanation is motivated by analyzing the step length during the adjustment of commitments. Figure 3.3 shows the mean distance between adjacent commitments sets in the

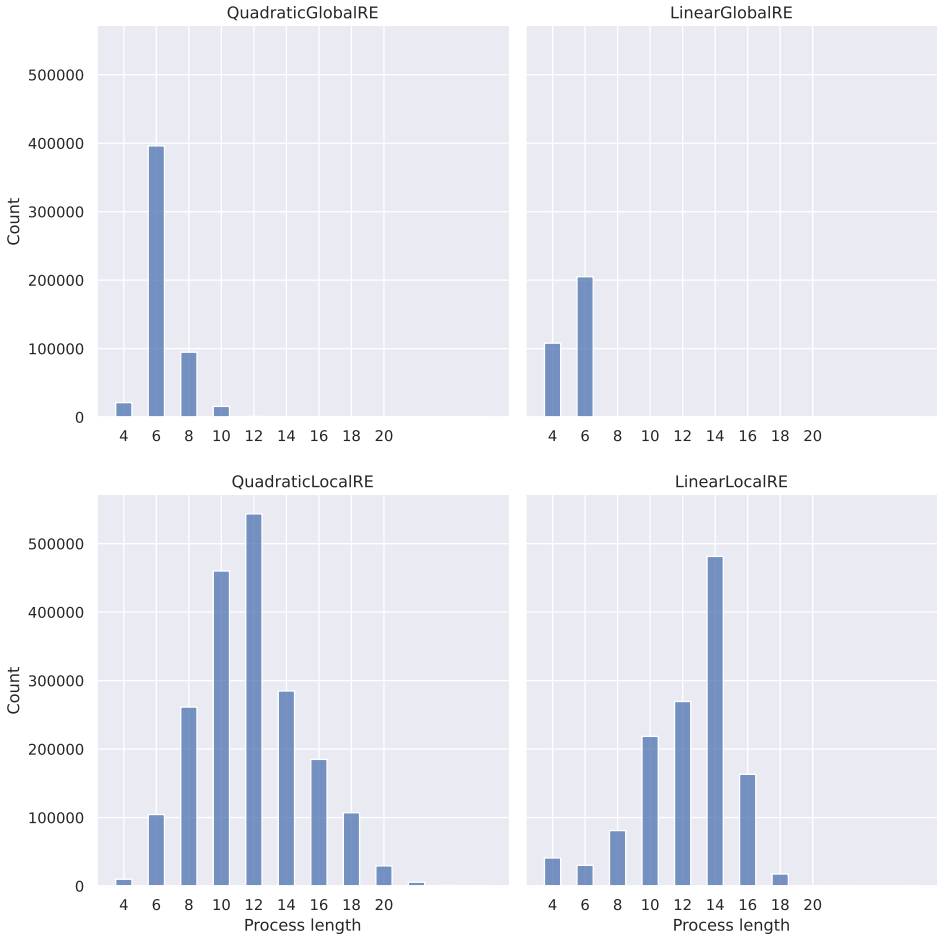


Figure 3.1: Distribution of process lengths for different models.

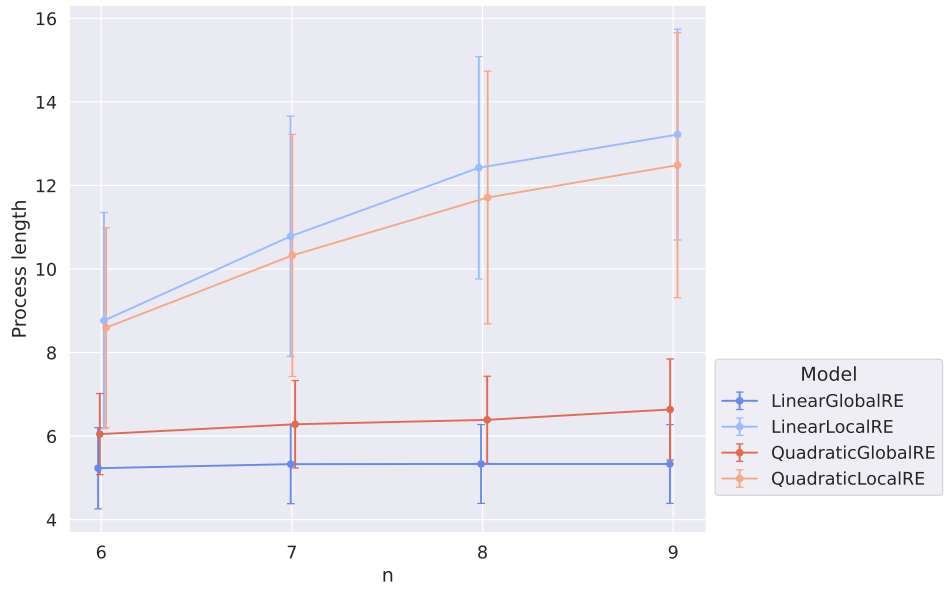


Figure 3.2: Mean process length for different models and sentence pools.

evolution of epistemic states over all branches. For simplicity, we measure the distance between two commitment sets by their simple Hamming distance, defined as the number of sentences not shared by both sets. For example, the simple Hamming distance between the commitments sets $\{s_1, s_2\}$ and $\{s_2, s_3\}$ is 2 since there are two sentences (s_1 and s_3) not shared by both sets.

Unsurprisingly, the locally optimizing models have roughly a mean step length of 1 since they are confined in their choice of new commitments to the 1-neighbourhood.[2] In contrast, the semi-globally optimizing models take bigger leaps with an increasing sentence pool size. Figure 3.4 shows why: With the increasing size of the sentence pool, the mean distance between initial commitments and fixed-point commitments increases. In other words, RE processes must overcome larger distances to reach their final states. Semi-globally optimizing models can walk this distance with fewer steps (Figure 3.2) since they can take comparably large steps (Figure 3.3). Locally optimizing models are confined to small steps (Figure 3.3) and, thus, have to take more steps (Figure 3.2).

[2]The mean distance is, for some cases, slightly greater than 1, which can be simply explained: The definition of the 1-neighbourhood is based on another Hamming distance than the one used here. In particular, there are sentence sets in the 1-neighbourhood of a sentence set whose simple Hamming distance is greater than 1. For instance, the set $\mathcal{C}_1 = \{s_1, \neg s_2\}$ is in the 1-neighbourhood of the sentence set $\mathcal{C}_2 = \{s_1, s_2\}$ since it only needs an attitude change towards one sentence (i.e., an attitude change towards s_2 from rejection to acceptance). However, the simple Hamming distance is 2 since both s_2 and $\neg s_2$ are not shared by $\mathcal{C}_1$ and $\mathcal{C}_2$.

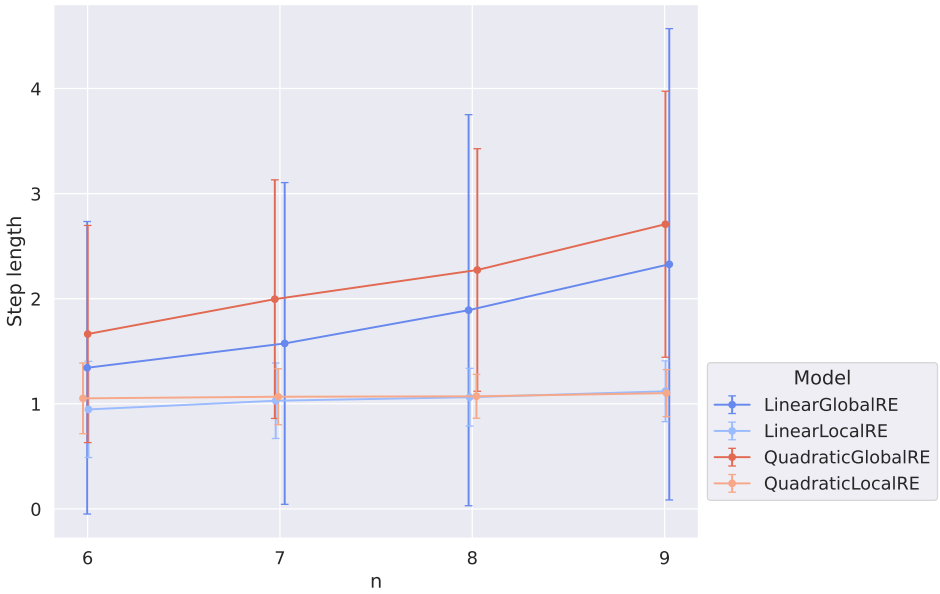


Figure 3.3: Mean step length of adjacent commitments for different models and sentence pools.

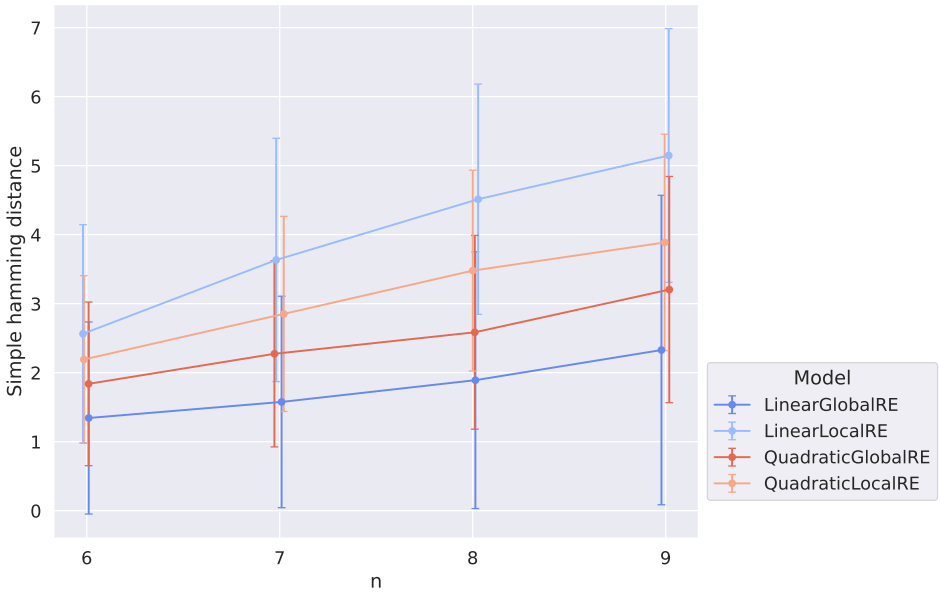


Figure 3.4: Mean distance between initial commitments and fixed points.

3.2 Global Optima

Global optima are fully determined by the achievement function of the RE model. Accordingly, global optima might differ between the linear and quadratic model variants but do not depend on whether the RE process is based on a local or semi-global optimization. In the following, we will therefore summarize analysis results with respect to global optima for linear models under the heading `LinearRE` and for quadratic models under the heading `QuadraticRE`.[3]

The mean number of global optima does not differ significantly between linear and quadratic models (5 ± 26 vs. 5 ± 14) and does not depend on the size of the sentence pool (see Figure 3.5).

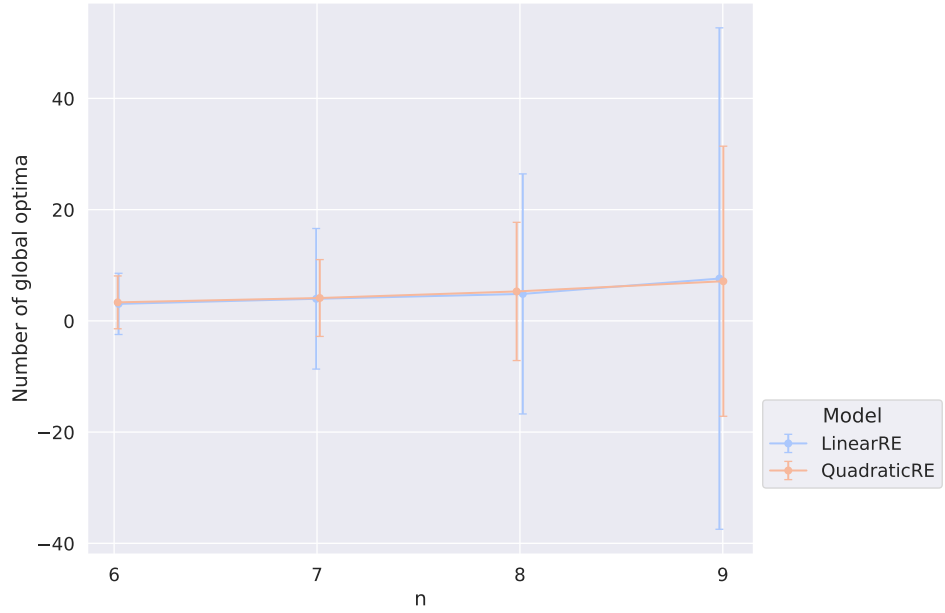


Figure 3.5: Number of global optima for different n.

However, the heatmap in Figure 3.6 shows an interesting dependence on the α-weights.

Here and in the following chapters, we will often rely on such heatmaps. Let us therefore provide some clarifications of their interpretation. If we are interested in visualising the dependence on α-weight configurations (i.e., a specific triples of α_A, α_F and α_S), it is sufficient to use two dimensions (α_A and α_S in our case) since the three weights α_A, α_F and α_S are not independent. The diagonals in these heatmaps from southwest to northeast are isolines for the faithfulness weight (α_F). In the following, we will refer to specific cells in these heatmaps in the typical

[3] In our data set, the analysis results might differ between semi-globally and locally optimizing models, which is, however, an artifact of the difference in interrupted model runs (i.e., model runs that could not properly end (see Section 2.4)). For the subsequent analysis of global optima, we rely on the model results of `QuadraticGlobalRE` and `LinearGlobalRE` since they had fewer interrupted model runs.

(x, y) fashion. For instance, we will call the cell with $\alpha_S = 0.5$ and $\alpha_A = 0.2$ the $(0.5, 0.2)$ cell.

Now, let's come back to Figure 3.6. For each simulation setup there is not necessarily one global optimum. Instead, there can be multiple global optima. Each cell in the heatmap provides for a specific α-weight configuration the mean number of global optima (over all simulation setups with this α-weight configuration). For the quadratic models, the number of global optima (and its variance) increases with an increase in α_S. For the linear models, on the other hand, the number of global optima is comparably low $(1 - 3)$ in all cells with the exception of the three islands $(0.4, 0.3)$, $(0.6, 0.2)$ and $(0.8, 0.1)$. These cells are characterised by $\alpha_F = \alpha_A$. For linear models, there are more ties in the achievement function under these conditions (see Appendix A), which results in an increase in global optima.

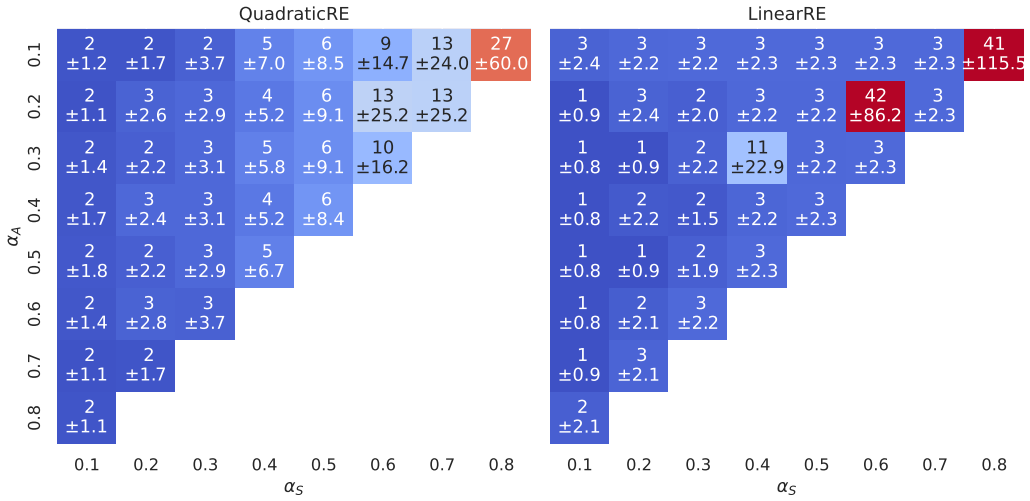


Figure 3.6: Mean number of global optima for different α-weight configurations.

Besides analysing the number of global optima, it is helpful to get a preliminary grasp on some topological properties of global optima. How are the commitments of global optima distributed over the space of all minimally consistent commitments? Are they located in a dense way to each other, or are they widely distributed in the whole space? What is their distance from initial commitments?

Figure 3.7 and Figure 3.8 depict the mean distance of global-optimum commitments in dependence of the sentence pool's size and α_F. We calculated for each configuration setup that has more than one global optimum the mean (simple Hamming) distance between global-optimum commitments and took the average of these means with respect to different ensemble properties. The share of configuration setups that have more than one global optimum is 0.58 over all models, 0.54 for linear models and 0.62 for quadratic models.[4]

[4]Note that global optima a process-independent. Hence, semi-globally and locally optimizing models do not differ with respect to their global optima.

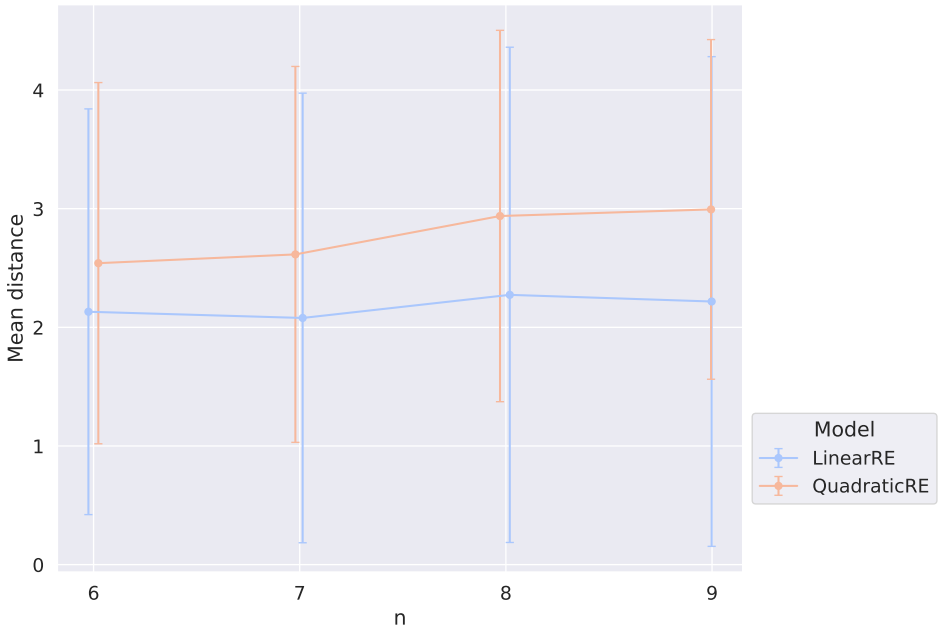


Figure 3.7: Mean distance of global-optima commitments for different n.

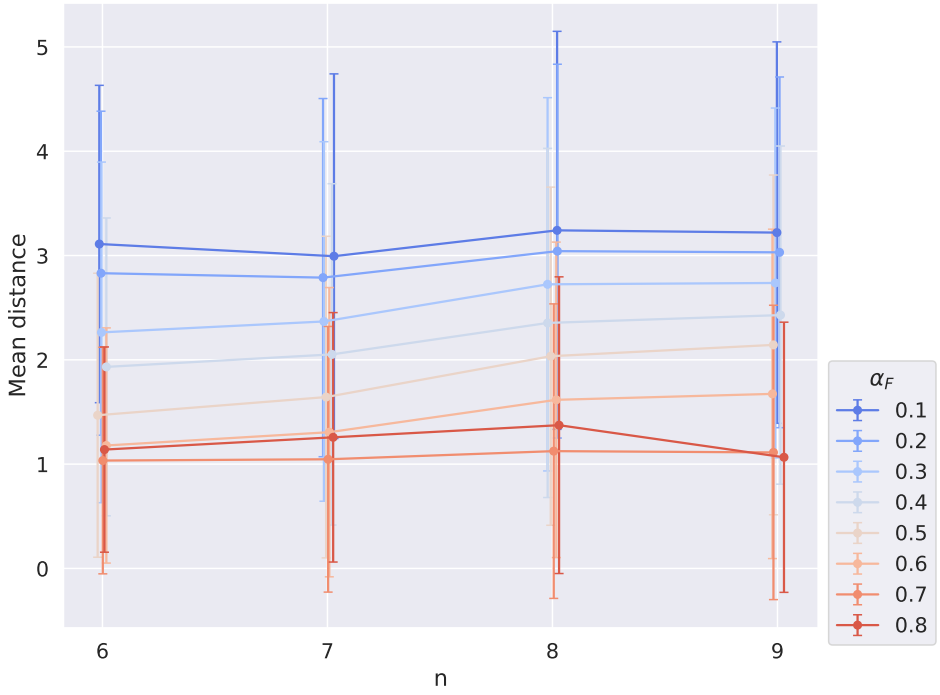


Figure 3.8: Mean distance of global-optima commitments for different α.

Figure 3.9 and Figure 3.10, one the other hand, depict the mean distance between initial commitments and global-optimum commitments. For that, we calculated for each simulation setup the mean (simple Hamming) distance between initial commitments and all global-optimum commitments of the simulation setup and, again, took the average of these means with respect to different ensemble properties.

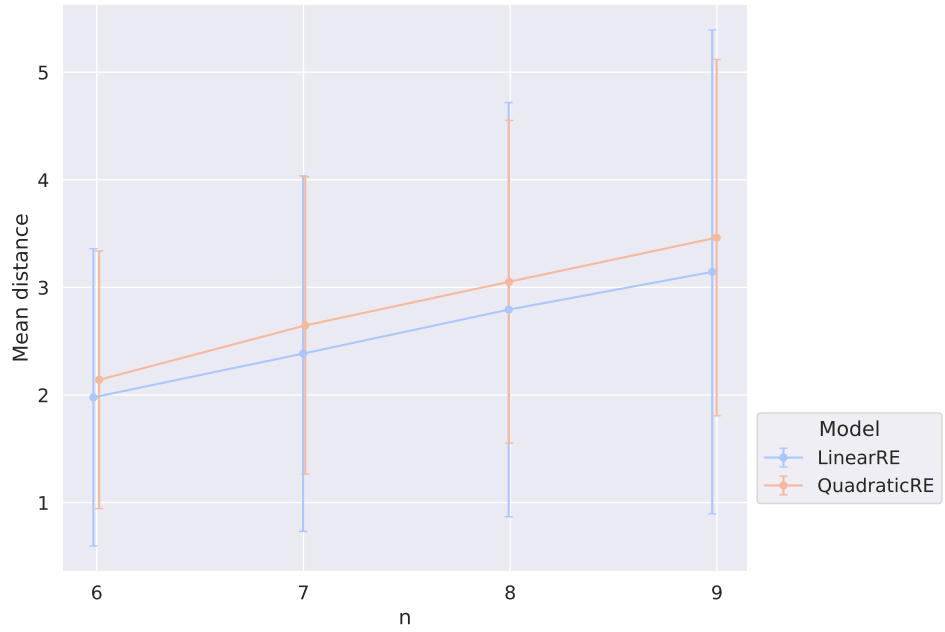


Figure 3.9: Mean distance between initial commitments and optimal commitments for different n.

Figure 3.7 and Figure 3.9 are hard to interpret. The mean distance of global optima does not seem to depend on the size of the sentence pool; the mean distance of initial commitments and global-optimum commitments might increase with the size of the sentence pool. However, without an additional consideration of larger sentence pools, we cannot draw these conclusions with certainty due to the large variance.

Figure 3.8 and Figure 3.10, one the other hand, show that the mean distance of initial commitments and global-optimum commitments as well as the mean distance between global-optimum commitments depend on α_F. The smaller α_F, the larger the distance. This result is not suprising. The weight α_F determines the extent to what final commitments should resemble initial commitments. You can think of α_F as the magnitude of an attractive force that pulls the commitments of the epistemic state to the initial commitments. Accordingly, if α_F gets smaller, global optima and fixed points will be distributed more widerspread in the space of epistemic states.

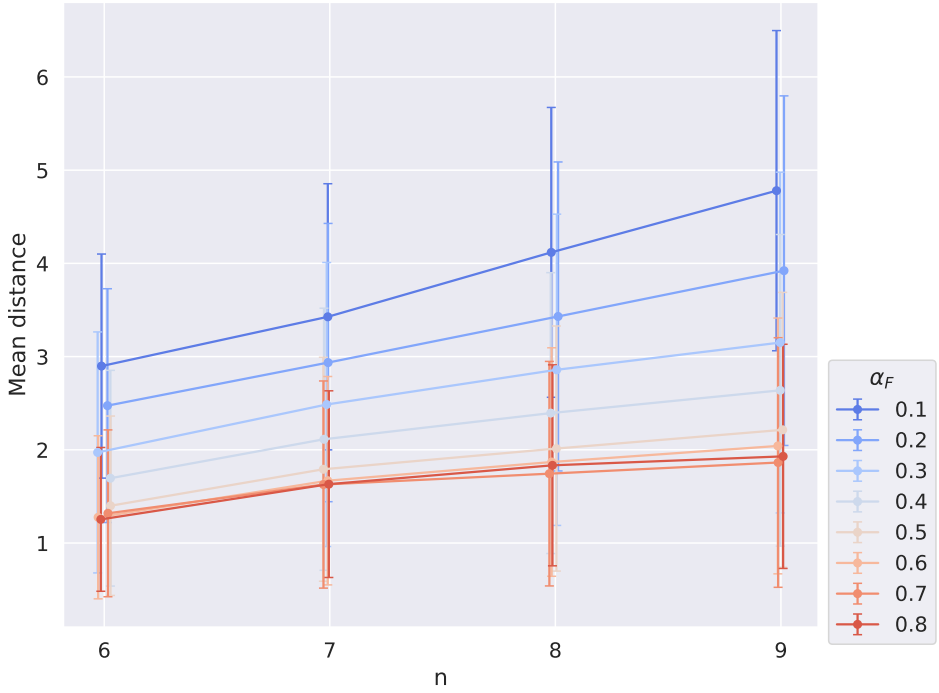


Figure 3.10: Mean distance between initial commitments and optimal commitments for different α.

3.3 Branching

The choice of a new theory (or a new set of commitments respectively) is underdetermined if there are different candidate theories (or commitment sets) that maximize the achievement of the accordingly adjusted epistemic state. In such a case, the model randomly chooses the new epistemic state. The model we use is able to track all these different branches to assess the degree of this type of underdetermination and to determine all possible fixed points for each configuration setup.

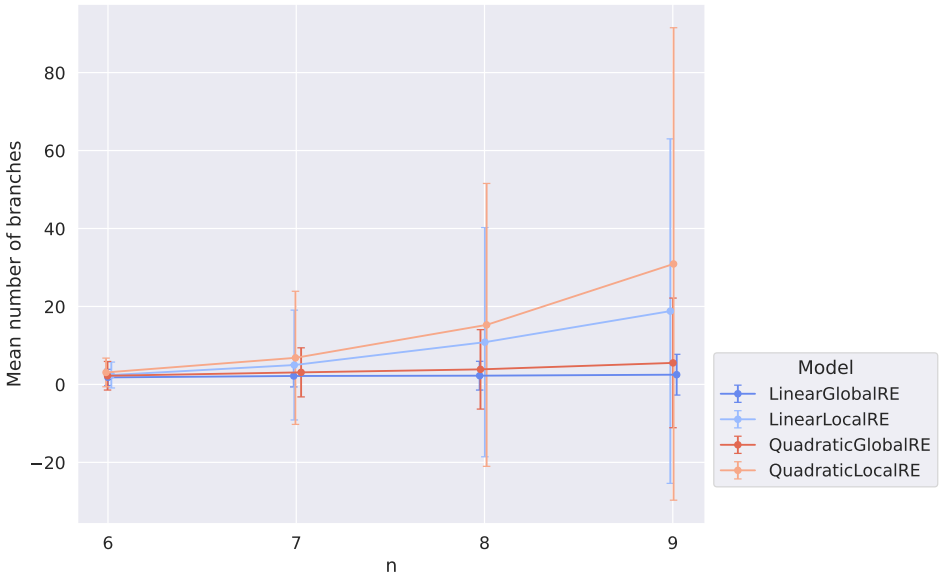


Figure 3.11: Mean number of branches for different models and sentence pools.

Figure 3.11 shows the mean number of branches with their dependence on the model and sentence pool. It suggests that branching is more prevalent in locally optimizing models. The large variance can be partly explained by the heat maps in Figure 3.12, which depict mean values (and standard deviations) for different weight combinations.

For `LinearGlobalRE` there are, again, islands with many branches (the cells $(0.4, 0.3)$, $(0.6, 0.2)$ and $(0.8, 0.1)$) which are characterised by $\alpha_F = \alpha_A$. The high number of branches correlates with a high number of fixed points (compare Figure 3.13) and a high number of global optima within these cells (compare Figure 3.6). We might, therefore, hypothesize that the model produces a high number of branches in these cells due to the high number of global optima.[5]

Interestingly, the identified hotspots of branches (and fixed points) for the `LinearGlobalRE` model are not reproduced by its locally optimizing cousin. This suggests that the

[5] In Chapter 4, we will analyze to what extent the model is able to reach these global optima. The numbers (7/8/8 branches and fixed points and 11/32/25 global optima) suggest that the number of fixed points are nevertheless not enough to reach all these global optima (see, e.g., Figure 4.6 and Figure 4.14 in Chapter 4).

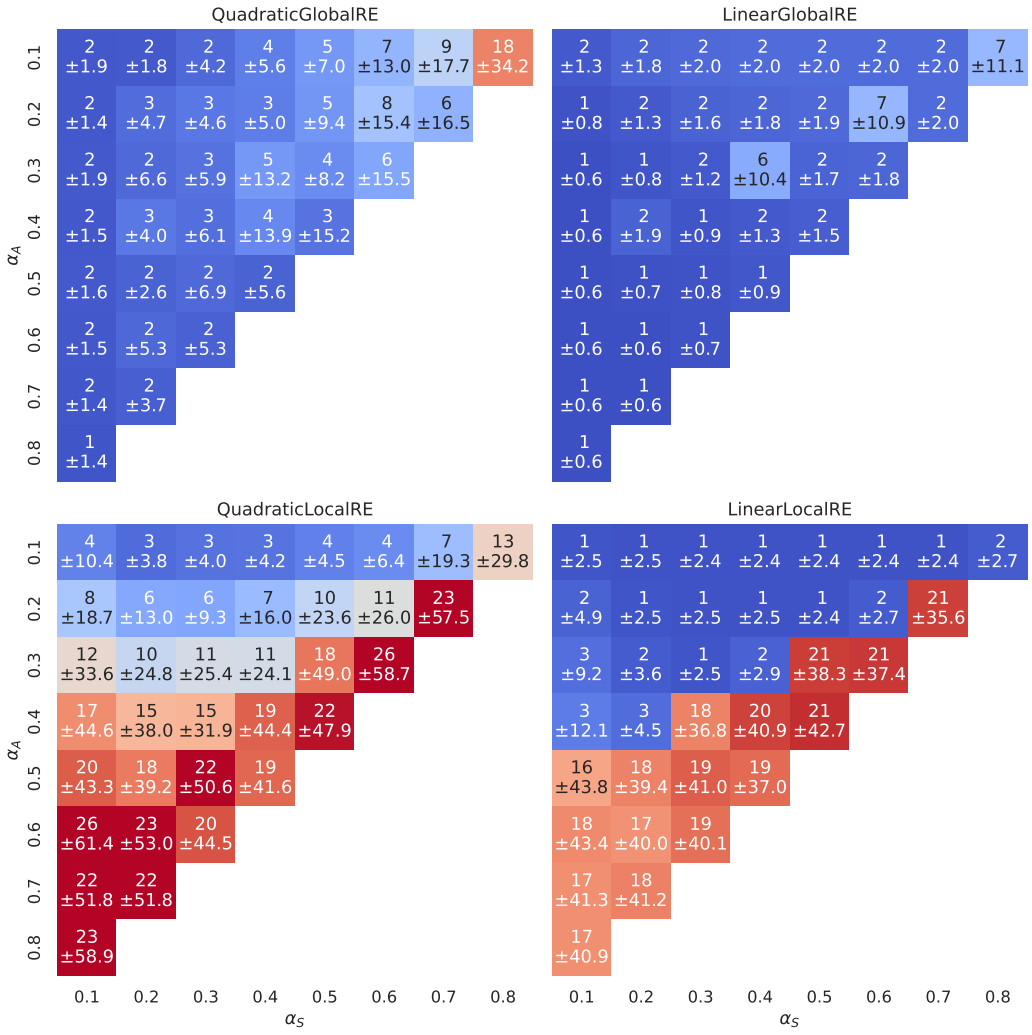


Figure 3.12: Mean number of branches for different models and weights.

`LinearLocalRE` model will perform worse than the `LinearGlobalRE` model to reach the increased amount of global optima.[6]

The "$\alpha_F = \alpha_A$"-line is, however, also relevant for the `LinearLocalRE` model. Above that line, branching is comparably low (roughly $1-3$) and below that line comparably high (with a high variance). The high number of branches does, however, not correlate with a high number of fixed points (see Figure 3.13). In other words, a lot of these branches end up in the same fixed point. This behaviour is to some extent even observable in the `QuadraticLocalRE` model.

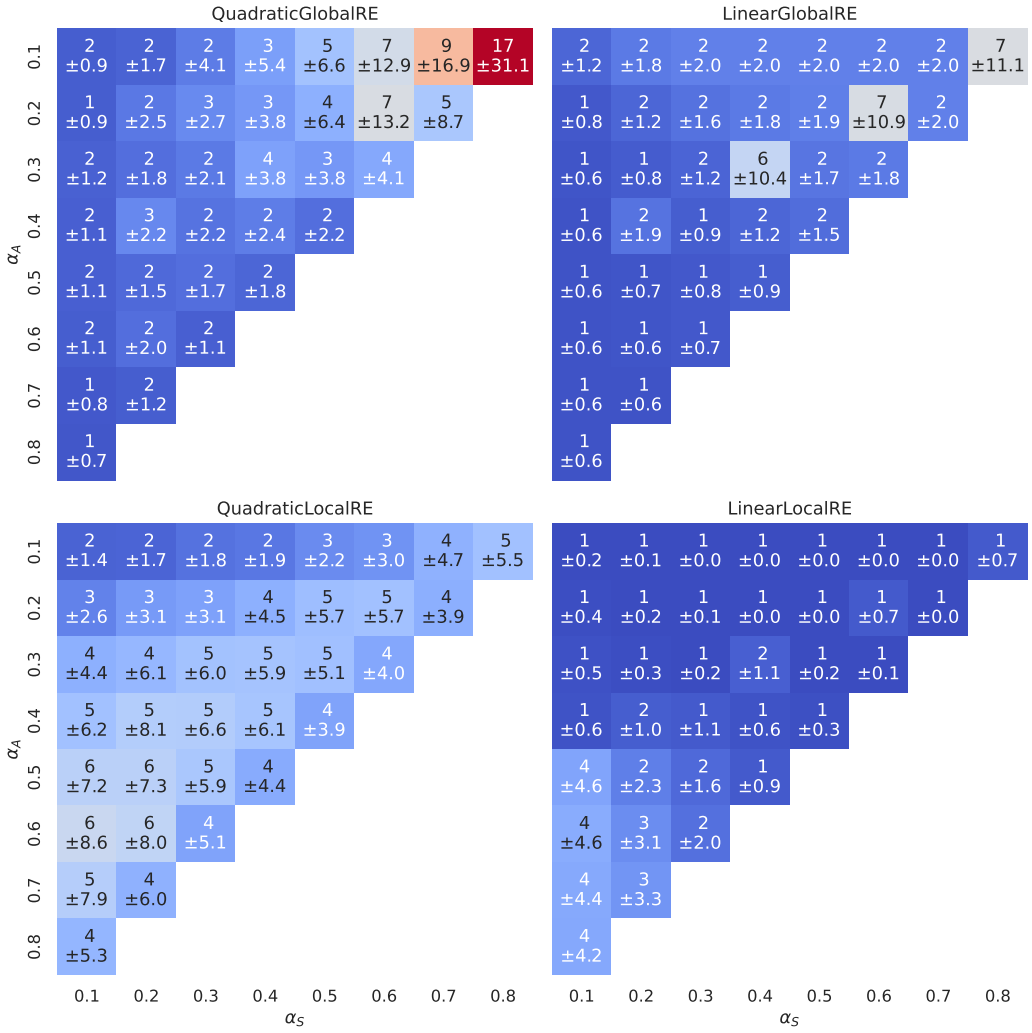


Figure 3.13: Mean number of fixed points for different models and weights.

[6] A hypothesis we will scrutinize in Chapter 4 (see, e.g., Figure 4.6 and Figure 4.14).

4 Global Optima and Fixed Points

4.1 Background

Global optima are epistemic states (i.e., commitments-theory pairs) that maximize the achievement function (see Chapter 2). The models we assess simulate RE processes by mutually adjusting commitments and theories. Since these models proceed in a semi-globally or locally optimizing fashion, fixed points of RE processes are not necessarily global optima (see Section 2.3 for details). It is, therefore, important to assess the performance of the different models with respect to their ability to reach global optima. Two main questions guide the following evaluation:

1. **GO efficiency**: Are fixed points global optima? More specifically, what is the share of global optima among fixed points?
2. **GO reachability**: Are global optima reachable by RE processes? More specifically, what is the share of fixed points among global optima?

GO efficiency and reachability might not only differ between model variants but might, additionally, depend on the specifics of the simulation setups. In the following, we will confine the consideration to the following dimensions:

- How do GO efficiency and reachability depend on the size of the sentence pool?
- How do GO efficiency and reachability depend on the arguments' mean number of premises?
- How do GO efficiency and reachability depend on α-weights?

We will answer these questions by calculating different relative shares in the following way.

Let the *ensemble* E be the entirety of simulation setups we used to simulate RE processes. Each simulation setup $e \in E$ corresponds to a set of RE processes that can evolve with this specific setup. Remember that the different steps in the evolution can be underdetermined. In other words, an RE process might branch. We will denote the set of all branches of a specific simulation setup e with B_e. Consequently, a specific setup can have more than one fixed point. Similarly, there is not necessarily one global optimum for each simulation setup but possibly many (denoted by GO_e).

GO efficiency can be calculated in two different ways. First, we can assess the share of global optima among all branches. In other words, we count those branches in B_e that end up in

global optima and divide by $|B_e|$. We will refer to this type of GO efficiency as *GO efficiency from the process perspective*. However, different branches might end up in the same fixed points. Another way of calculating GO efficiency—*GO efficiency from the result perspective*—avoids a possible "multiple" counting of fixed points by considering the *(mathematical) set* of fixed points.

More formally, let $\{FPGO\}_e$ be the set of all fixed points of e that are global optima, and let $[FPGO]_e$ be the fixed points of all branches in e that are global optima. The latter is formally a multiset, which can contain one fixed point multiple times. We can now define different types of GO efficiency—one based on $\{FPGO\}_e$ and one on $[FPGO]_e$. The corresponding share will be calculated by formulas of the form

$$GOE^{proc}(E^*) := \frac{\sum_{e \in E^*} |[FPGO]_e|}{\sum_{e \in E^*} |B_e|}$$

and of the form

$$GOE^{res}(E^*) := \frac{\sum_{e \in E^*} |\{FPGO\}_e|}{\sum_{e \in E^*} |\{FP\}_e|}$$

with respect to different subsets $E^* \subset E$.

For instance, let E_{M_1} be the set of all simulation setups belonging to the model M_1. We can calculate the overall GO efficiency of M_1 from the process perspective by $GOE^{proc}(E_{M_1})$ and from the result perspective by $GOE^{res}(E_{M_1})$.

How can we interpret these different types of GO efficiency? One idea is to interpret them probabilistically. According to this suggestion, the ensemble-based model assessment informs us about the probabilities of catching global optima by means of RE processes. On this view, GO efficiency from the process perspective is the probability of a process ending up in a global optimum. On the other hand, GO efficiency, from the result perspective, is the probability of a fixed point being a global optimum. You can think of the difference in terms of when or under which conditions to ask about the probability. In contrast to the latter case, you do not know the fixed point of the process (perhaps the process has not ended yet) in the former case.

It does not make much sense to distinguish GO reachability between the process and result perspective. GO reachability asks about the share of global optima that are reachable by RE processes among all global optima. Naturally, the denominator is the (mathematical) set of all global optima in a simulation setup (GO_e), which is a process-independent property of the simulation setup. Since it might happen that $|[FPGO]_e| > |GO_e|$ we should define GO reachability based on $\{FPGO\}_e$:

$$GOR_{E^*} := \frac{\sum_{e \in E^*} |\{FPGO\}_e|}{\sum_{e \in E^*} |GO_e|}$$

4.2 Results

> **Note**
>
> The results of this chapter can be reproduced with the following Jupyter notebook: https://github.com/re-models/re-technical-report/blob/main/notebooks/data_analysis_chapter-go-and-fp.ipynb.

4.2.1 Model Overview

Table 4.1 and Figure 4.1 provide an overview of the different models' overall GO efficiency and reachability.

Model	GO efficiency (result perspective)	GO efficiency (process perspective)	GO reachability
LinearGlobalRE	0.73	0.73	0.33
LinearLocalRE	0.45	0.54	0.14
QuadraticGlobalRE	0.76	0.75	0.49
QuadraticLocalRE	0.33	0.35	0.27

Table 4.1: Overall GO efficiency and reachability of the different models

The semi-globally optimizing models perform better than the locally optimizing models regarding all measures.

GO efficiency is high for the former ($0.73 - 0.76$) and does not differ (much) between the process and result perspective. For locally optimizing models, GO efficiency varies between 0.33 and 0.54. We only observe a difference between the process and result perspective for the `LinearLocalRE` model (0.45 vs. 0.54). In other words, the extent of branching for the `LinearLocalRE` model differs between those processes that end up in global optima and those which do not.

GO reachability is below GO efficiency for all models and varies between low (0.14 for `LinearLocalRE`) and medium (0.49 for `QuadraticGlobalRE`).

With respect to the overall GO efficiency and reachability, the `QuadraticGlobalRE` model performs best since it reaches the highest value in GO reachability and is slightly better than `LinearGlobalRE` concerning GO efficiency.

For the locally optimizing models, the comparison between quadratic and linear shaped G functions is less clear-cut: While `LinearLocalRE` performs better in GO efficiency than

`QuadraticLocalRE` (0.45/0.54 vs. 0.33/0.35), it is the other way around concerning GO reachability (0.14 vs. 0.27).

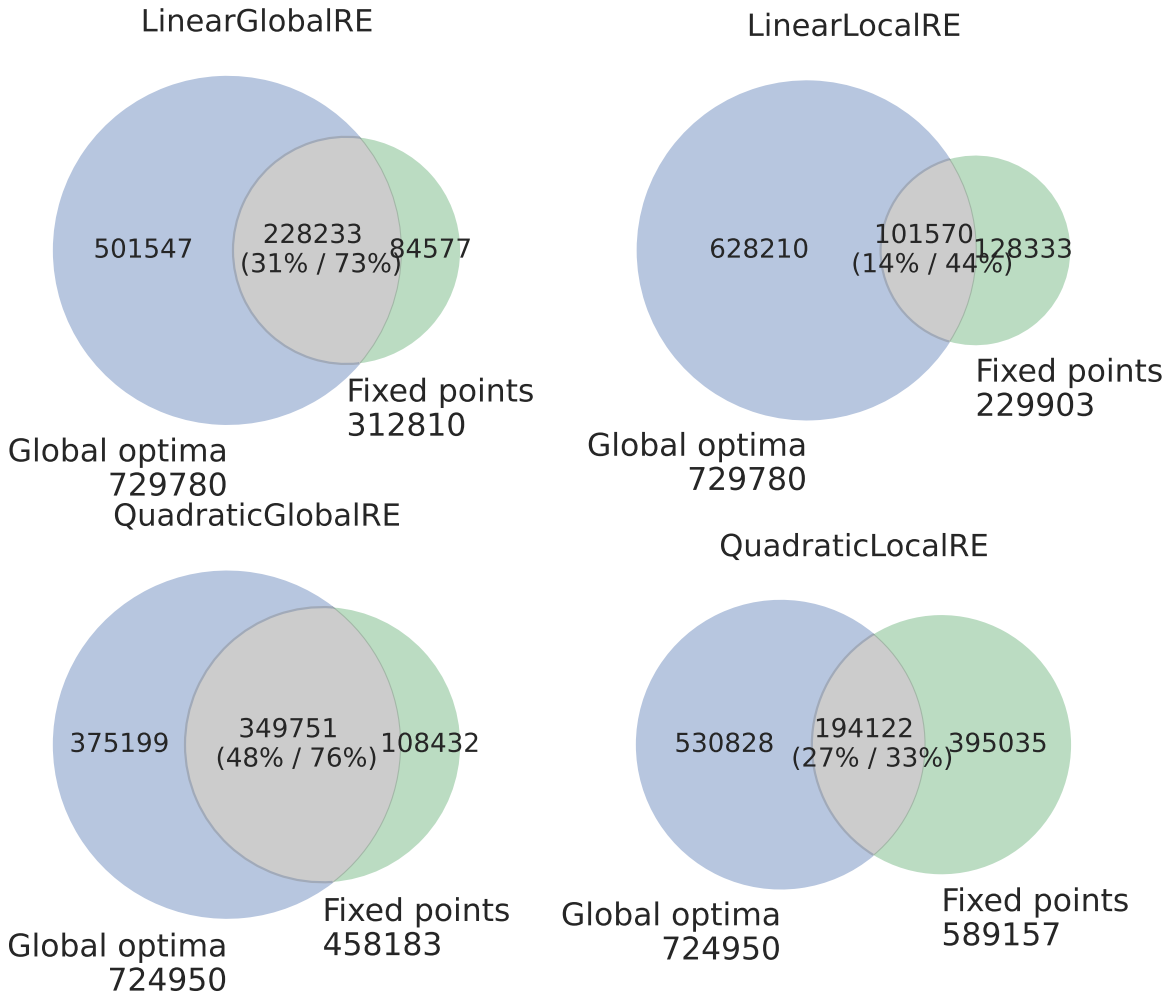


Figure 4.1: Overall GO efficiency (result perspective) and reachability of the different models.

4.2.2 GO Efficiency

4.2.2.1 Dependence on Sentence Pool

Figure 4.2 shows that GO efficiency is more or less stable along different sizes of the sentence pool for semi-globally optimizing models. The locally optimizing models not only perform worse than the semi-globally optimizing, but GO efficiency decreases for them with an increase in the size of the sentence pool.

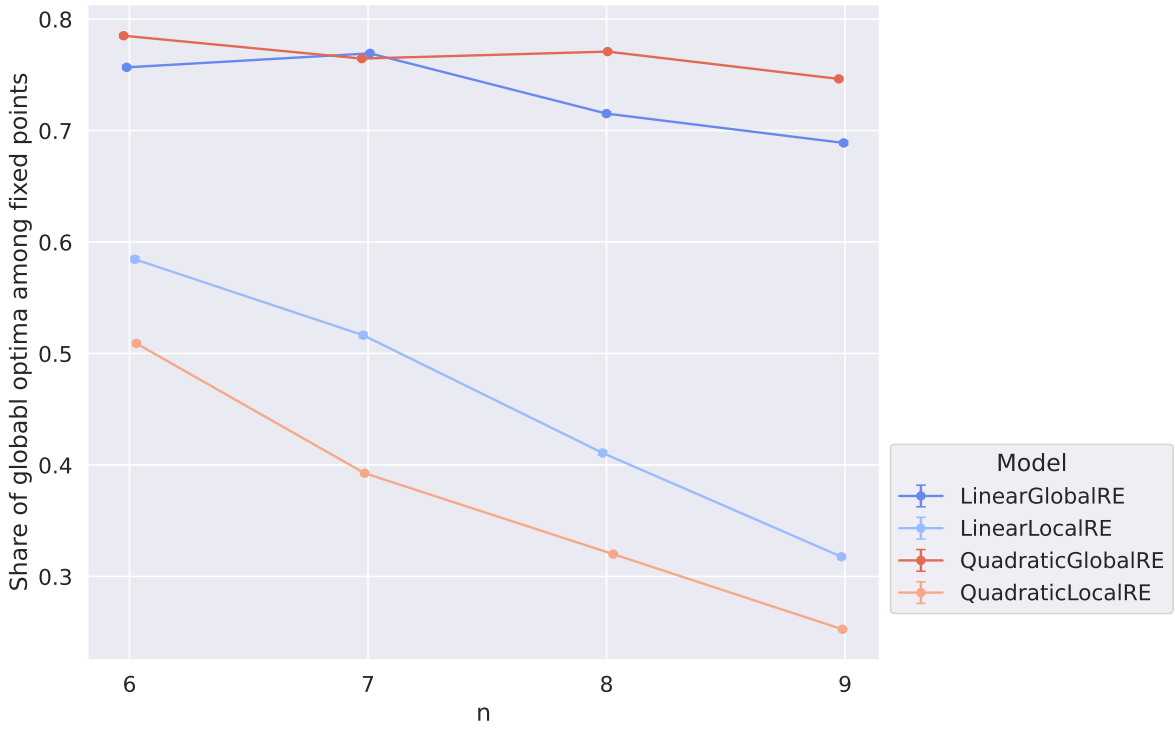


Figure 4.2: Dependence of GO efficiency (result perspective) on the size ($2n$) of the sentence pool.

As we already saw in the model overview, there is no big difference between the result and process perspective except for the `LinearLocalRE` model, which performs better from the process than from the result perspective (see Figure 4.3).

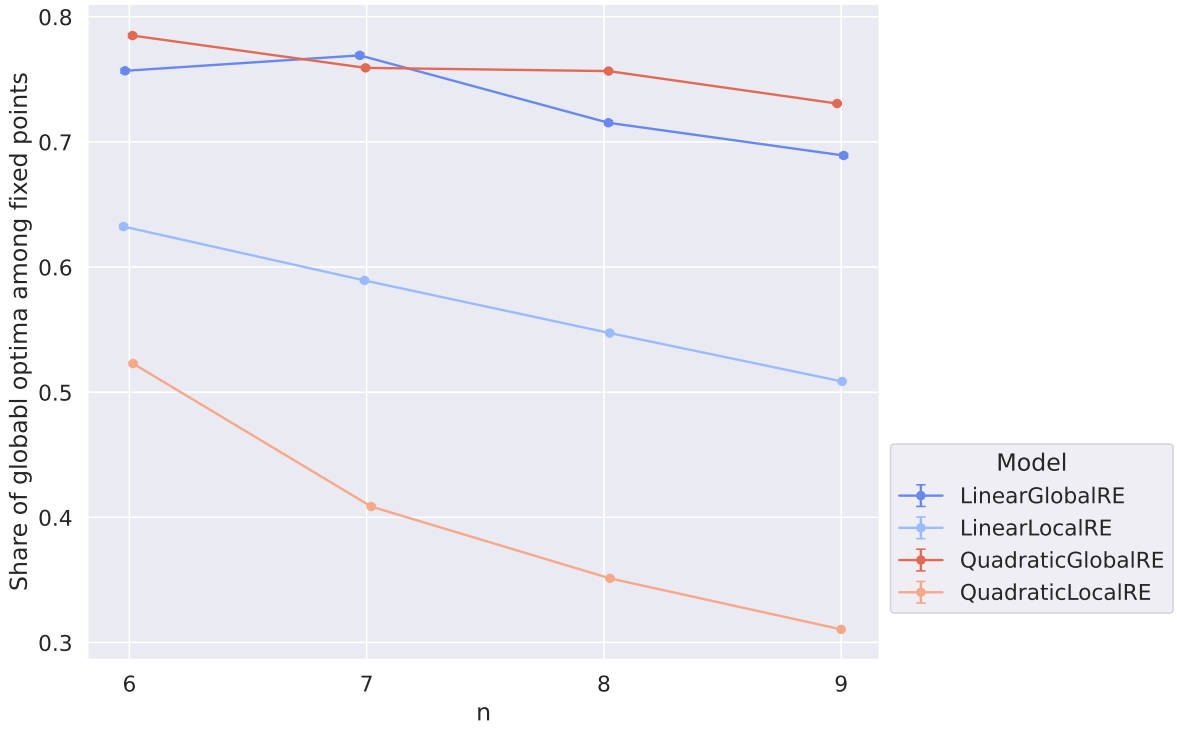


Figure 4.3: Dependence of GO efficiency (process perspective) on the size ($2n$) of the sentence pool.

4.2.2.2 Dependence on Mean Number of Premises

Figure 4.4 and Figure 4.5 show the dependence of GO efficiency on the mean number of the arguments' premises. They might be interpreted as suggesting that the locally optimizing models tend to perform worse with an increasing amount of premises in arguments. At least the difference between semi-globally and locally optimizing models is smaller for lower mean numbers of premises.

The zigzag shape of the lines suggests that the actual underlying variance is bigger than the pictured error bars.[1] One explanation might be that GO efficiency depends crucially on properties of the dialectical structures other than the mean number of premises. Since there are few dialectical structures for individual data points, their calculation is hardly based on a

[1]The error bars are standard deviations, which are calculated by bootstrapping on the used subset E^* in the calculation of $GOE(E^*)$.

representative sample. Accordingly, the zigzag might indicate the variation in GO efficiency more accurately. Consequently, the plots must be interpreted with caution.

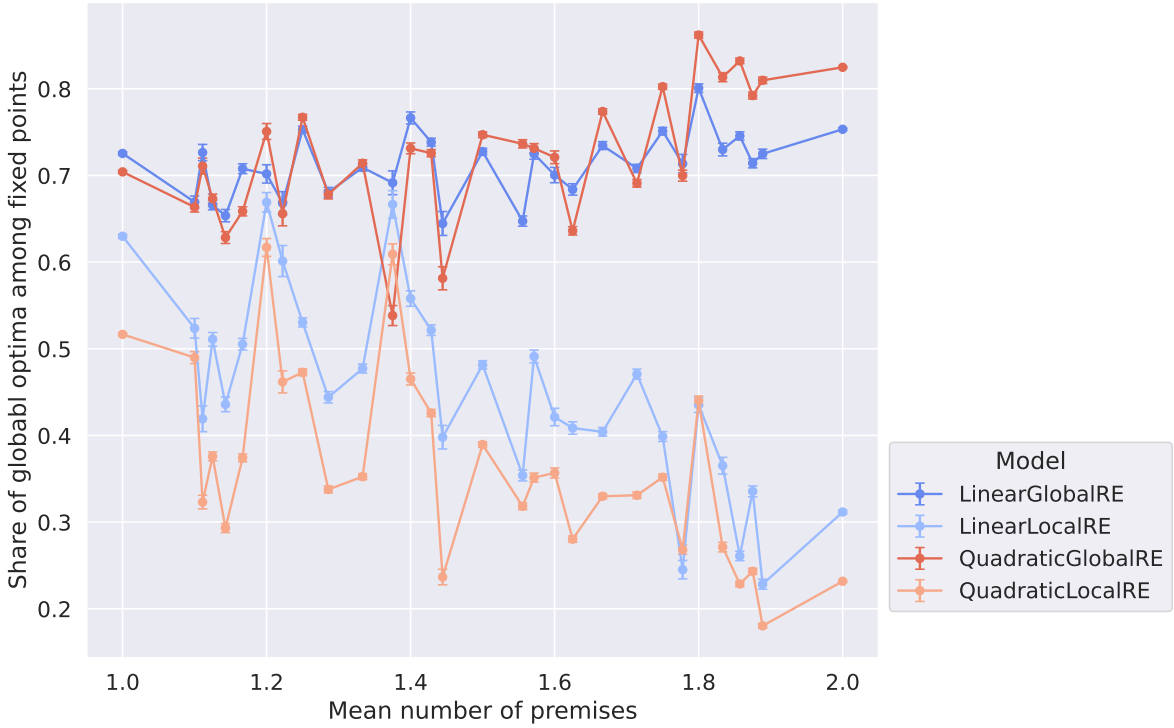


Figure 4.4: Dependence of GO efficiency (result perspective) on the mean number of arguments' premises.

4.2.2.3 Dependence on α-Weights

In the preceding sections, we aggregated over the spectrum of different α-weight configurations. The question is to what extent GO efficiency depends on the chosen α-weights.

The heatmaps in Figure 4.6 and Figure 4.7 provide an overview of the α-weight dependence. In the following, we will refer to specific cells in the typical (x, y) fashion. For instance, we will call the cell with $\alpha_S = 0.5$ and $\alpha_A = 0.2$ the $(0.5, 0.2)$ cell.

GO efficiency tends to increase with a decrease in α_A and with an increase in α_S. There are some exceptions to this pattern, especially in linear models. Most notably, there are four "cold" islands in the linear models from both perspectives (compare the $(0.2, 0.4)$, $(0.4, 0.3)$, $(0.6, 0.2)$ and $(0.8, 0.1)$ cells in Figure 4.6 and Figure 4.7). The comparably diminished magnitude of GO efficiency can be explained by the comparably high number of global optima in three of theses cells (compare the $(0.4, 0.3)$, $(0.6, 0.2)$ and $(0.8, 0.1)$ cells in Figure 3.6). Surprisingly, the

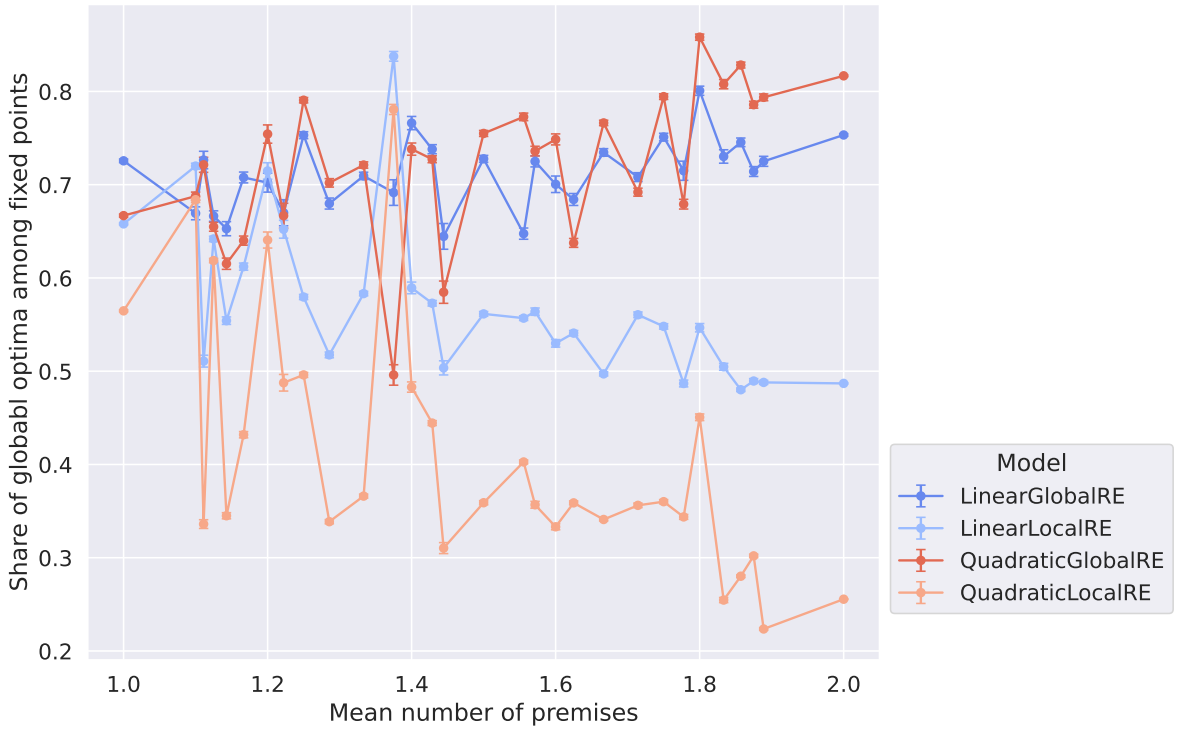


Figure 4.5: Dependence of GO efficiency (process perspective) on the mean number of arguments' premises.

locally and semi-globally optimizing model perform similarly bad, although the semi-globally optimizing model produces much more branches and fixed points in these cells (compare Figure 3.12 and Figure 3.13).

Additionally, linear models tend to exhibit more extreme values than quadratic models. In other words, the difference between "hot" and "cold" regions is higher for linear models than for the quadratic counterparts.

Figure 4.8 and Figure 4.9 can be used to compare semi-globally with locally optimizing models. For each α cell, they show the difference in GO efficiency between the semi-globally optimizing model and its locally optimizing variant. As already observed above, the locally optimizing models perform on average worse than the semi-globally optimizing models. The difference in performance is smaller between the linear variants than the quadratic variants. The `LinearLocalRE` model is for some α-weight combinations even better than the `LinearGlobalRE` and for many configurations as good as the latter.

Figure 4.10 and Figure 4.11 show, additionally, the dependence on the mean number of arguments. The mean number of premises varies between 1 and 2. We divided this interval into four bins ($1 - 1.25$, $1.25 - 1.5$, $1.5 - 1.75$ and $1.75 - 2$) and every heatmap row aggregates over those dialectical structures that have a mean number of premises in the corresponding bin.

Interestingly, there is a difference between quadratic and linear models. For the linear models, the heatmaps do not change much with an increase in the mean number of premises. However, heatmaps suggest such a dependence for the quadratic models: The higher the mean number of premises, the higher the difference between semi-globally and locally optimizing models.

4.2.3 GO Reachability

4.2.3.1 Dependence on Sentence Pool

Figure 4.12 shows that GO reachability drops quickly for the linear models and slightly for the quadratic ones with increasing size of the sentence pool. For $n = 9$, a locally optimizing model (`QuadraticLocalRE`) even outperforms a semi-globally optimizing model (`LinearGlobalRE`).

4.2.3.2 Dependence on Mean Number of Premises

As before, the overall performance in dependence on the mean number of premises is hard to interpret. Figure 4.13 might suggest that the three models `LinearGlobalRE`, `QuadraticLocalRE` and `LinearLocalRE` perform worse with an increase in the mean number of premises. Only `QuadraticGlobalRE` is able to keep its level of performance.

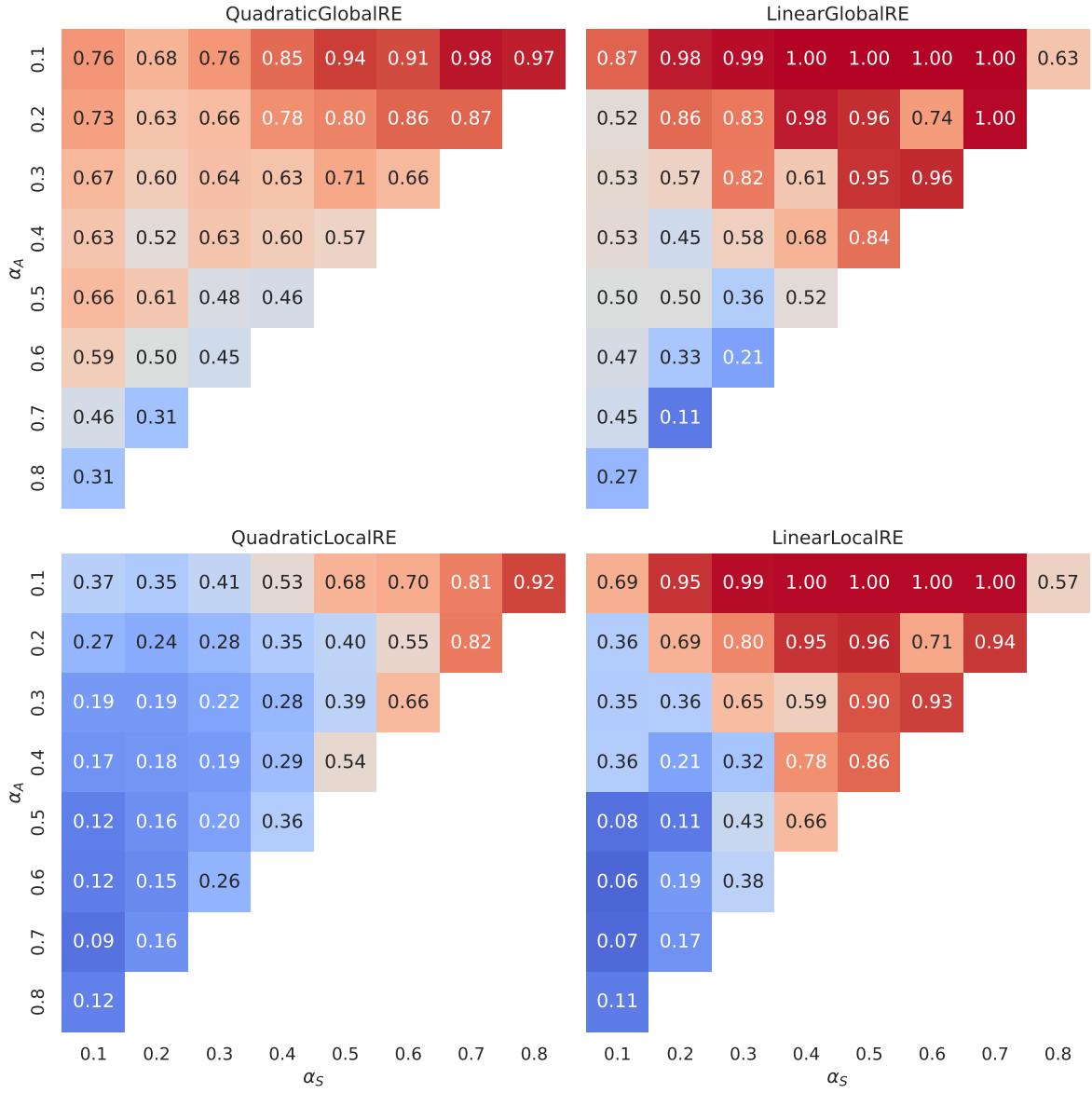


Figure 4.6: Dependence of GO efficiency (result perspective) from α-weights for the different model variants.

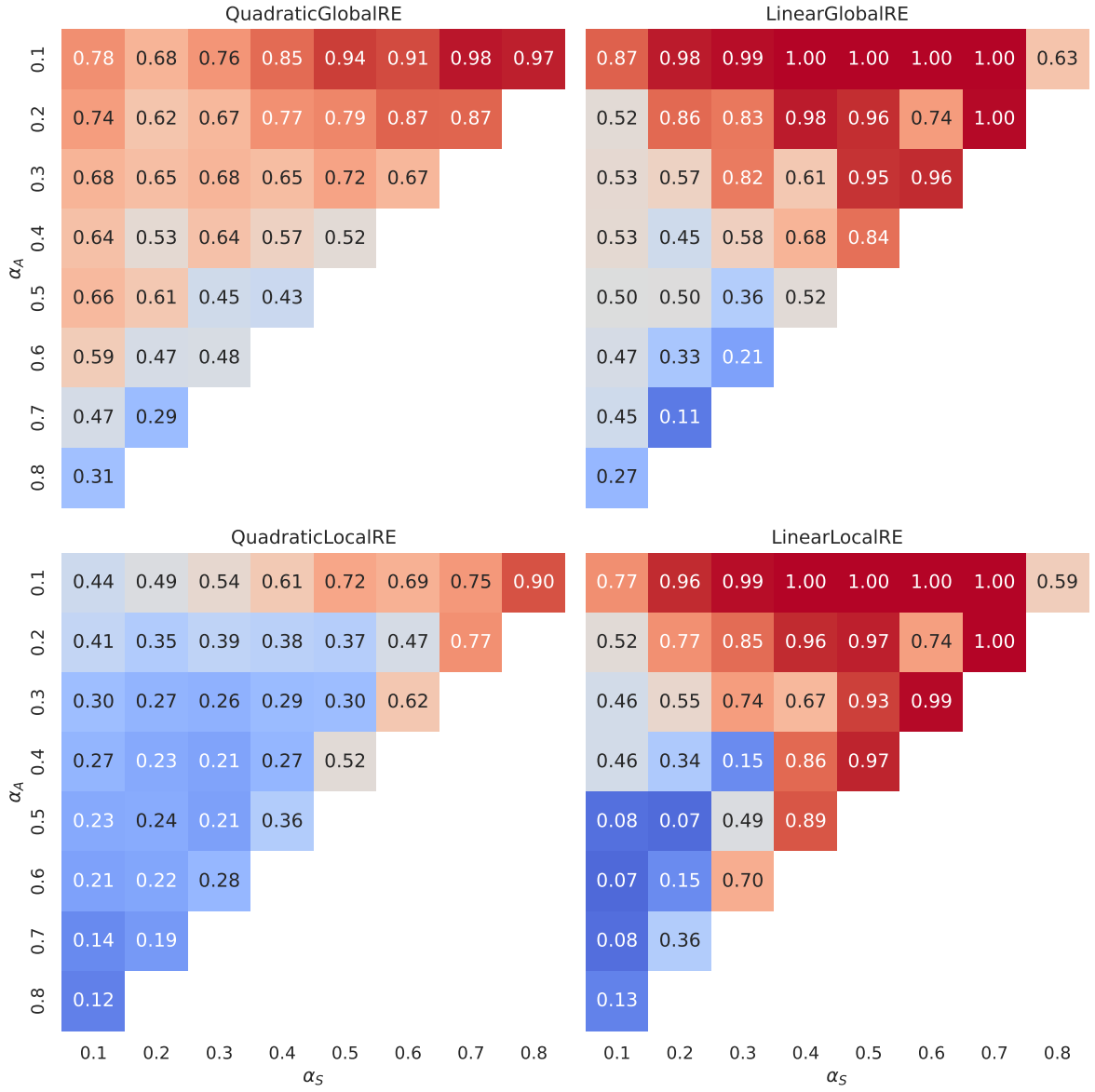


Figure 4.7: Dependence of GO efficiency (process perspective) from α-weights for the different model variants.

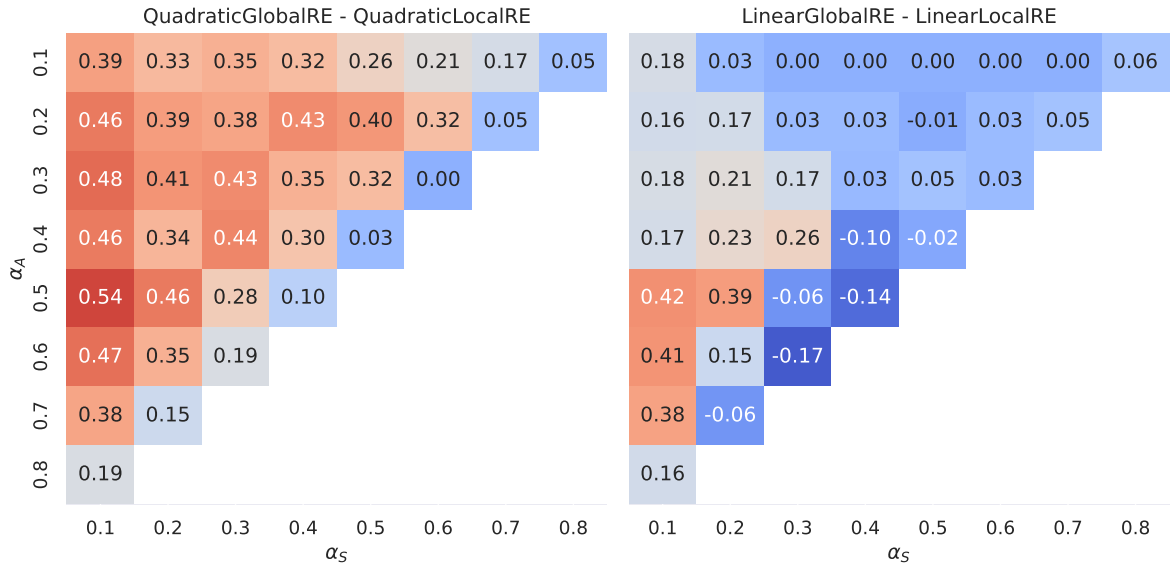


Figure 4.8: Comparing GO efficiency (result perspective) between semi-globally and locally optimizing models for different α-weights.

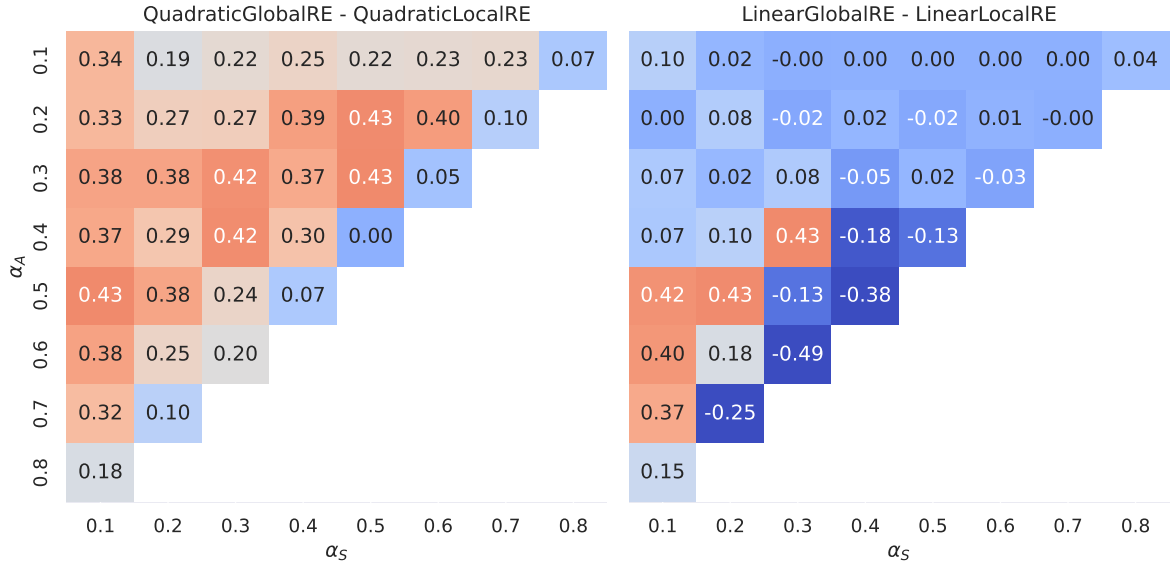


Figure 4.9: Comparing GO efficiency (process perspective) between semi-globally and locally optimizing models for different α-weights.

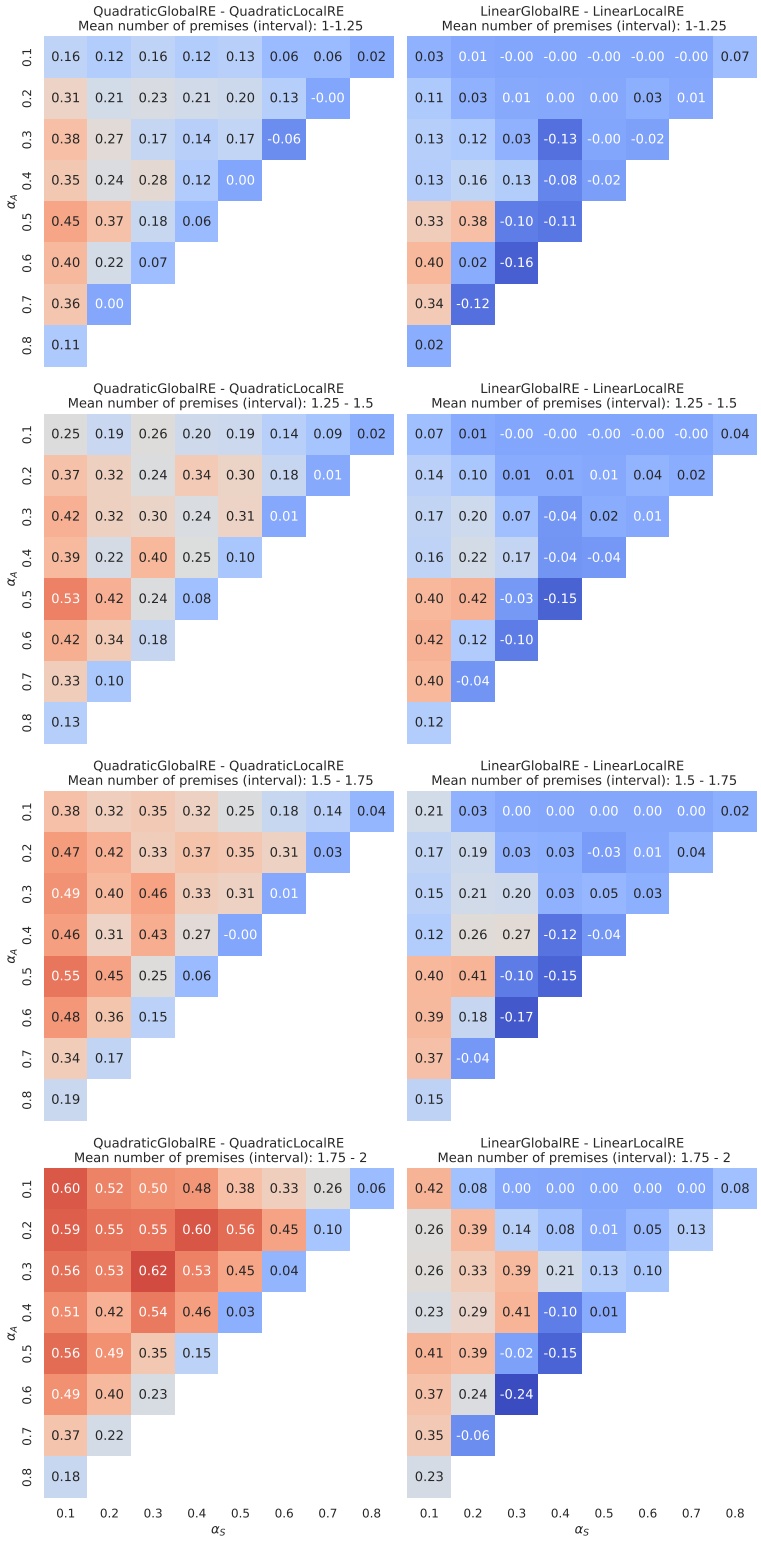


Figure 4.10: Comparing GO efficiency (result perspective) between semi-globally and locally optimizing models for different α-weights and intervals of the mean number of arguments' premises.

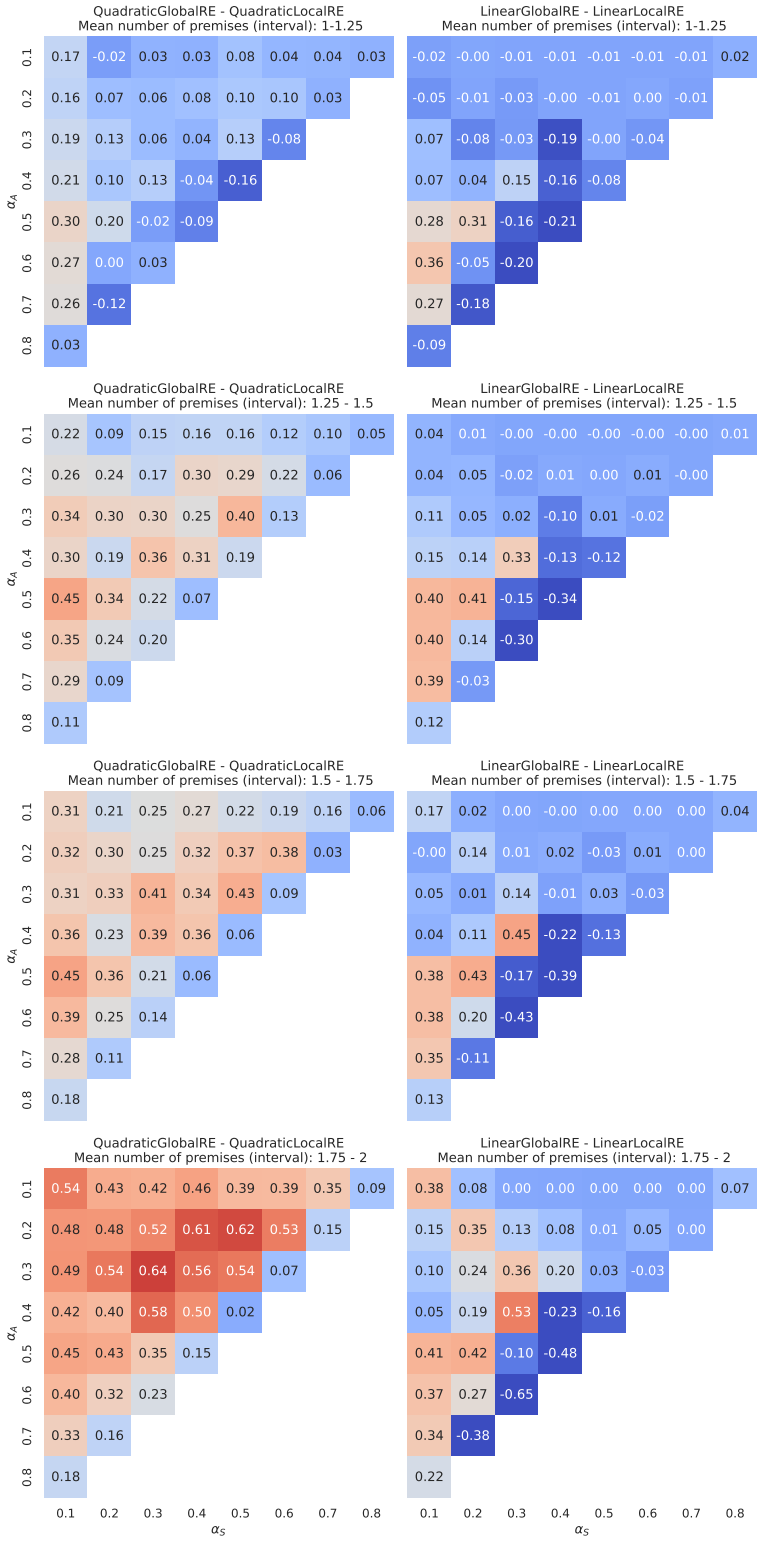


Figure 4.11: Comparing GO efficiency (process perspective) between semi-globally and locally optimizing models for different α-weights and different intervals of the mean number of arguments' premises. 50

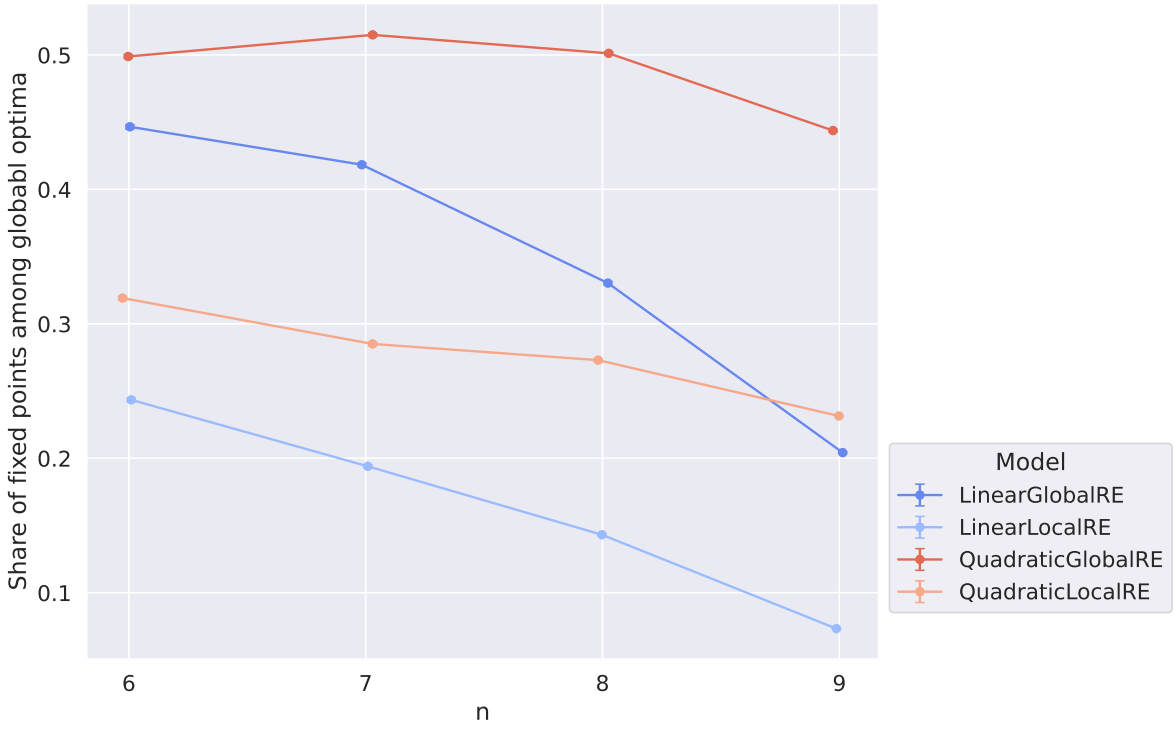


Figure 4.12: Dependence of GO reachability on the size ($2n$) of the sentence pool.

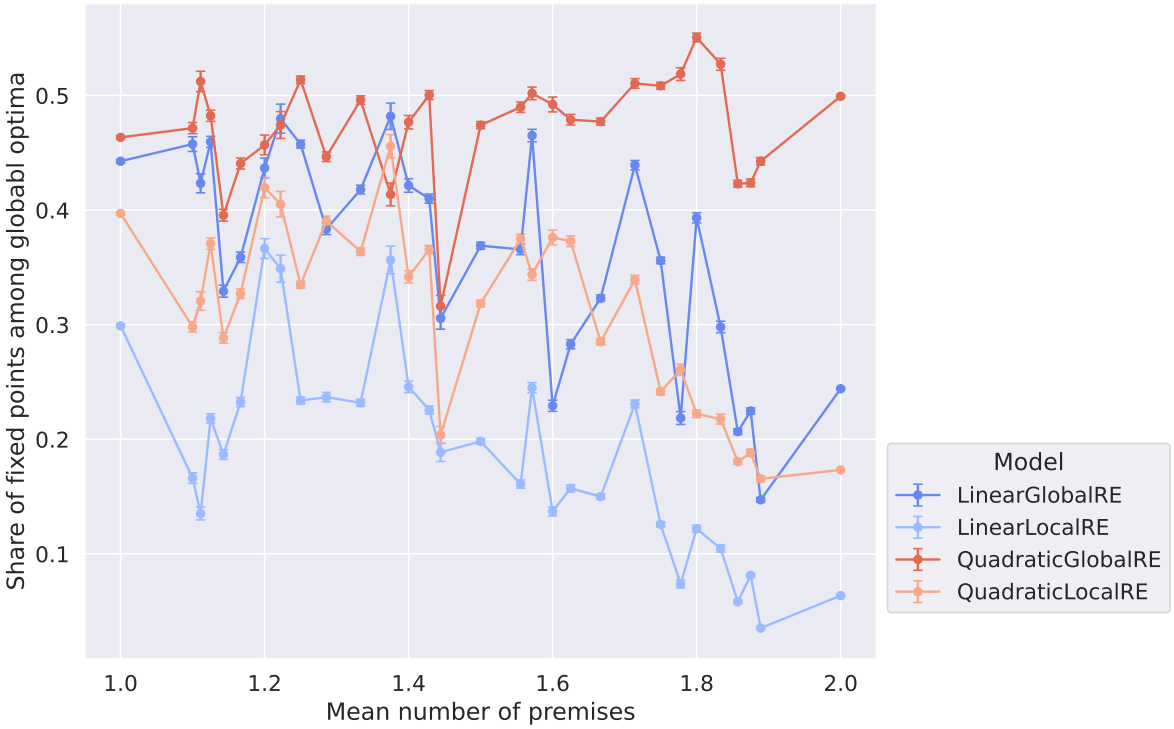


Figure 4.13: Dependence of GO reachability on the mean number of arguments' premises.

4.2.3.3 Dependence on α-Weights

The dependence of GO reachability on α-weights is somewhat similar to that of GO efficiency. For the semi-globally optimizing models, GO reachability tends to increase with a decrease in α_A and an increase in α_S. Again, there are exceptions to this behaviour. Besides the islands of the `LinearGlobalRE` model, the 0.1 α_F isoline has particularly low GO reachability values for the `QuadraticGlobalRE` model.

The linear model variants' cold islands can, again, be explained by the comparably high number of global optima in three of theses cells (compare the $(0.4, 0.3)$, $(0.6, 0.2)$ and $(0.8, 0.1)$ cells in Figure 3.6).

The locally optimizing model variants have a comparably non-regular dependence on α-weights. Additionally, the values do not vary that much between different cells as compared to the globally optimizing models.

The direct comparison between semi-globally and locally optimizing models (Figure 4.15) shows that locally optimizing models are, for some α-weight combinations, able to outperform the semi-globally optimizing models (cells with negative values).

By separating dialectical structures according to their mean number of premises (Figure 4.16) we can assess whether GO reachability depends on the mean number of premises: The advantage of semi-globally optimizing models as roughly indicated by the "hot" cells in the $(0.2-0.7, 0.1-0.2)$ area in Figure 4.15 increases with the mean number of premises. In contrast, the positions of cells for which locally optimizing models outperform semi-globally optimizing models (roughly, the "cold" cells of the 0.1/0.2 α_F isolines in Figure 4.15) do not depend that much on the mean number of premises.

4.3 Conclusion

On average, GO efficiency is high for semi-globally optimizing models and medium-high for locally optimizing models. The fact that for locally optimizing models GO efficiency drops with the size of the sentence pool is, to some extent, worrisome since they are intended to be used in scenarios with larger sentence pools, which are computationally too demanding for semi-globally optimizing models. The question is whether their performance can be improved by increasing their search depth d.

However, in specific contexts the modeller will choose a specific set of α-weights. We already saw that the performance of the different models varies significantly between different α-weight configurations. Consequently, the dependence on the sentence pool should be assessed for those regions of α-weight configurations that are of interest to the modeller. For instance, if we choose to confine the analysis to α-weight configurations with $\alpha_A < \alpha_S$, the `LinearLocalRE` model outperforms every other model in GO efficiency (see Figure 4.17).

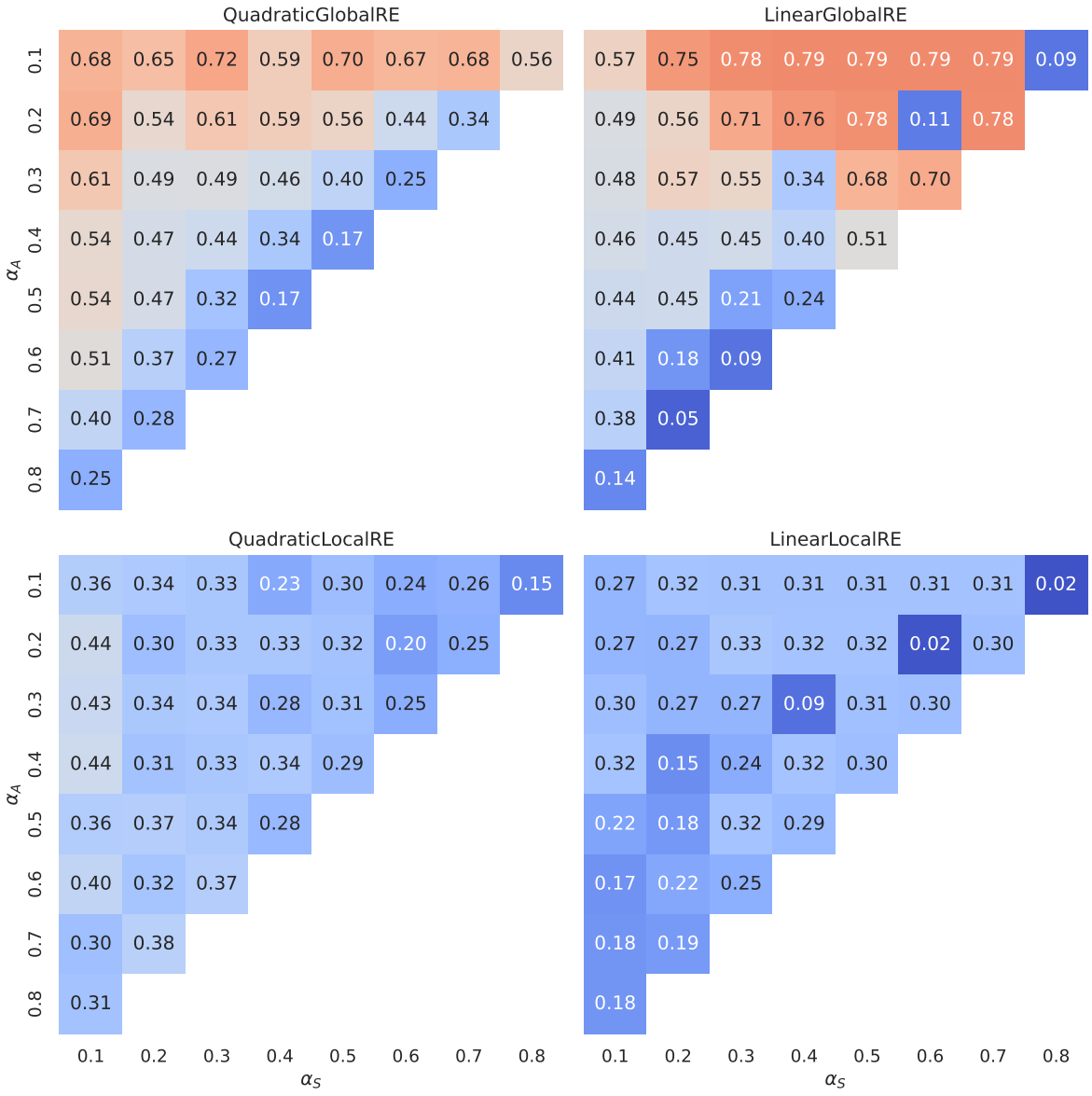


Figure 4.14: Dependence of GO reachability from α-weights for the different model variants.

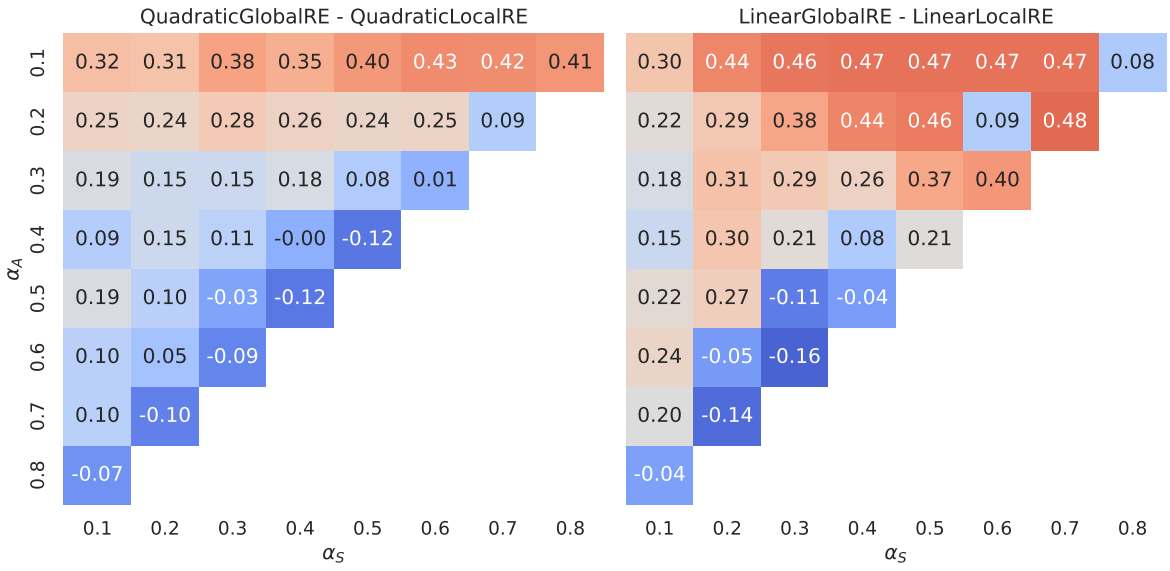


Figure 4.15: Comparing GO reachability between semi-globally and locally optimizing models for different α-weights.

Surprisingly, GO reachability is low to medium for all models. Additionally, all but the `QuadraticGlobalRE` model perform worse with an increase in the size of the sentence pool. A better understanding of this behaviour requires a more detailed analysis, which should be based on a more extensive set of dialectical structures.

The `QuadraticGlobalRE` model outperforms all other models on average. A direct comparison of the locally optimizing models is complicated since it involves a trade-off: While the `LinearLocalRE` model reaches a higher GO efficiency than the `QuadraticLocalRE` model, it is the other way around with respect to GO reachability.

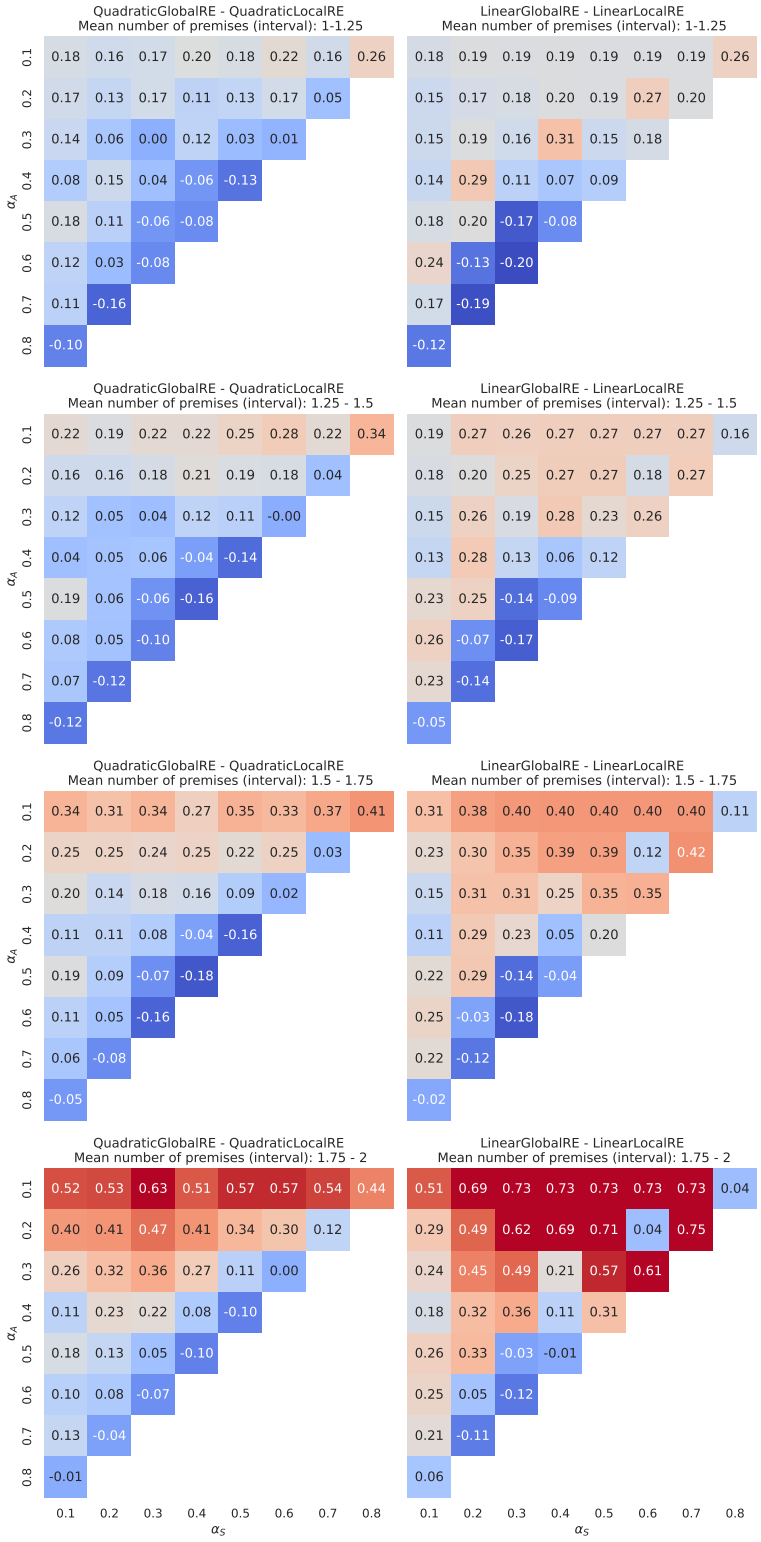


Figure 4.16: Comparing GO reachability between semi-globally and locally optimizing models for different α-weights and different intervalls of the mean number of arguments' premises.

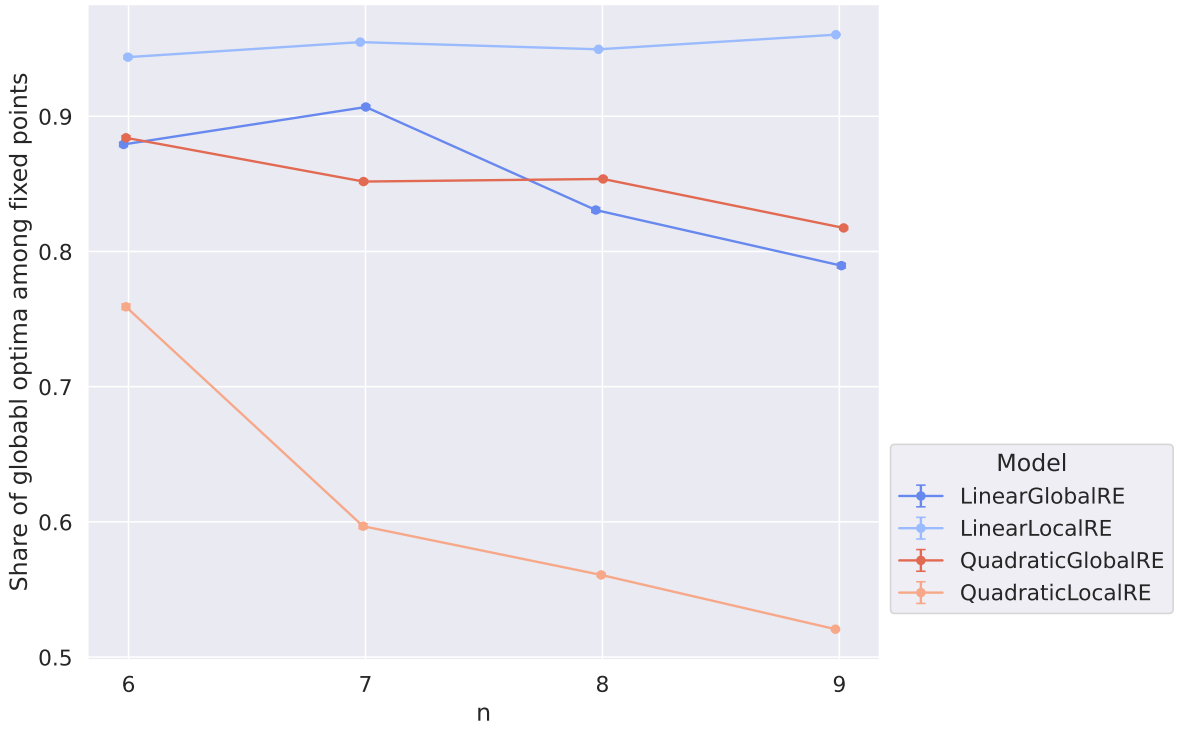


Figure 4.17: Dependence of GO efficiency (process perspective) on the size ($2n$) of the sentence pool for $\alpha_A < \alpha_S$.

5 Full RE States

5.1 Background

RE is commonly understood as an account of justification, and the aspired outcomes of applying RE are equilibrium states, which are supposed to be justified according to RE.

Consequently, it is interesting to study the formal counterparts in the model that represent, or at least approximate, equilibrium states: *full RE states*. A theory-commitment-pair $(\mathcal{C}, \mathcal{T})$ is a *full RE state* if and only if they live up to very high standards, namely,

1. if it is a global optimum according to the achievement function and
2. the theory $\mathcal{T}$ fully and exlusively accounts for the commitments $\mathcal{C}$.

The second criterion amounts to the requirement that every commitment and no other sentence of the sentence pool is derivable from the theory, given the arguments of the dialectical structure in the background.

An RE model is not required to yield a full RE state in every case. However, from the viewpoint of model evaluation, it may still desirable to have a model that is at least somewhat likely to reach full RE states. This is especially relevant to the fixed points of locally optimizing model variants, which have a severely restricted set of options at every adjustment step.

Still, whether the attainment of full RE states is important, will depend on the objectives pursued with a specific application of RE (or formal models thereof). If, for example, the objective is making up one's mind, gaining understanding of a subject matter, or if we take justification to come in degrees rather than being a yes-or-no matter, less than full RE states may be completely satisfactory outcomes.

Note that both fixed points and global optima can qualify as a full RE states. Hence, we present the results for global optima and fixed points separately. For the latter, we distinguish again between the result and the process perspective.

5.2 Results

> **Note**
>
> The results of this chapter can be reproduced with the following Jupyter notebook: https://github.com/debatelab/re-technical-report/blob/main/notebooks/data_analysis_chapter_full_re_states.ipynb.

5.2.1 Overall Results

5.2.1.1 Global Optima

Model	Relative share of full RE global optima	Number of full RE global optima	Number of global optima
QuadraticRE	0.115	82318	714584
LinearRE	0.275	192559	700830

Table 5.1: Relative share of full RE states among global optima

Observations

- The relative share of full RE states among global optima is substantially higher for linear model variants than for quadratic models (Figure 5.1).
- The small differences in Table 5.1 between semi-globally optimizing model variants and their globally optimizing counterparts are but an artifact of the model implementation. They can be explained by differences in interrupted model runs (see Section 3.2).

5.2.1.2 Fixed Points

Model	Relative share of full RE fixed points	Number of full RE fixed points	Number of fixed points
QuadraticGlobalRE	0.093	42660	458147
LinearGlobalRE	0.235	73477	312783
QuadraticLocalRE	0.052	30616	588236
LinearLocalRE	0.198	45241	228122

Table 5.2: Relative share of full RE states among fixed points (result perspective)

Observations

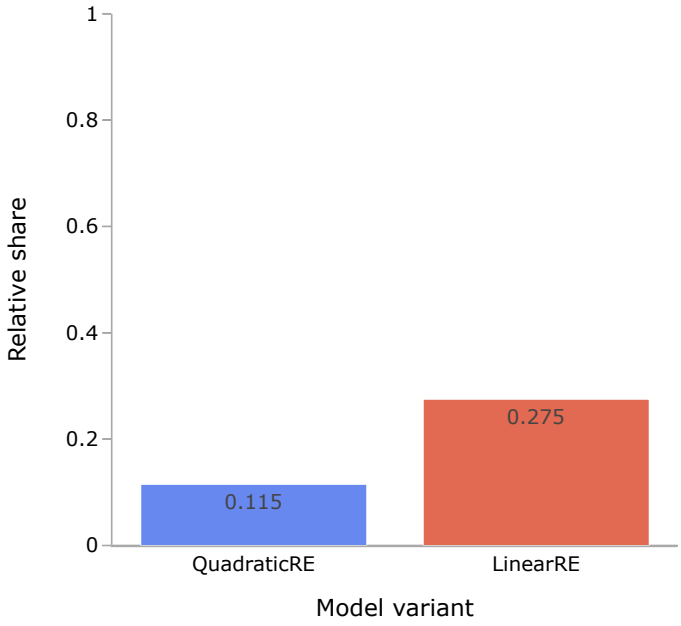


Figure 5.1: Relative share of full RE states among global optima grouped by model variant

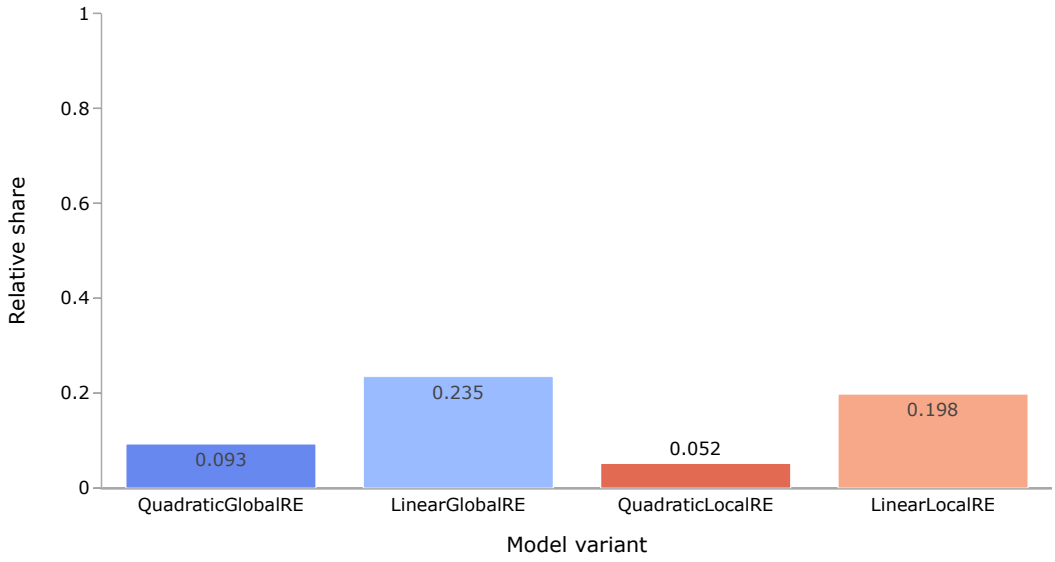


Figure 5.2: Relative share of full RE states among fixed points (result perspective) grouped by model variant

- The relative share of full RE fixed points from the result perspective (Figure 5.2) is lower than the corresponding results for global optima (Figure 5.1). This result is unsurprising as fixed points are reached through semi-globally or locally optimizing processes, which cover a restricted search space in contrast to global optimization.[1]
- From the result perspective, the relative shares of full RE fixed points of quadratic model variants are substantially lower than those of their corresponding linear model variants.

Model	Relative share of full RE fixed points	Number of full RE fixed points	Number of fixed points
QuadraticGlobalRE	0.088	46644	528616
LinearGlobalRE	0.235	73492	313002
QuadraticLocalRE	0.081	162044	1991852
LinearLocalRE	0.479	623825	1303077

Table 5.3: Relative share of full RE states among fixed points (process perspective)

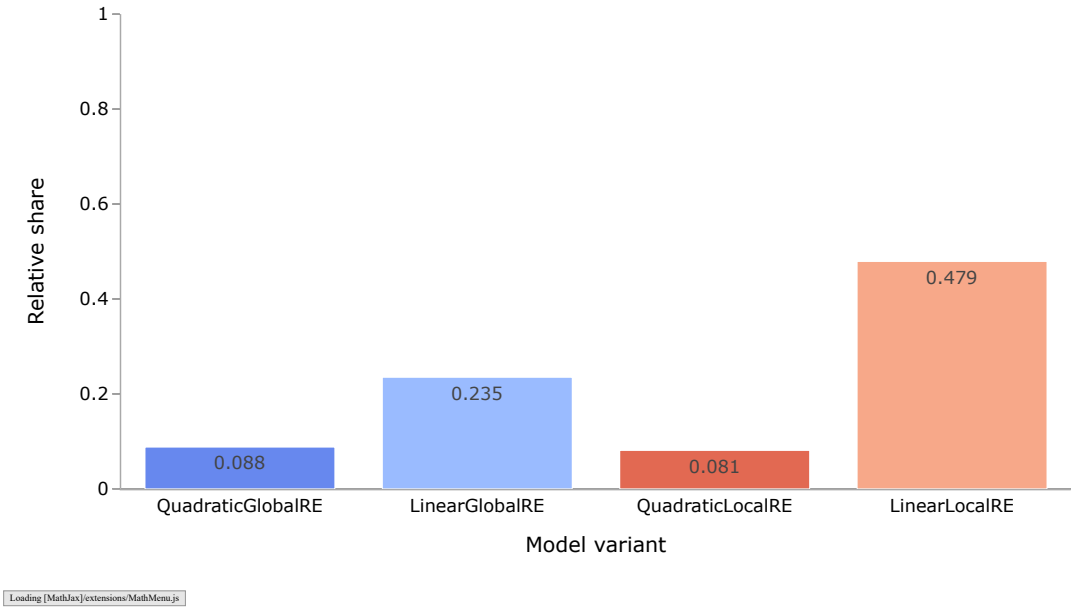


Figure 5.3: Relative share of full RE states among fixed points (process perspective) grouped by model variant

[1]For the difference between result and process perspective, see Section 4.1.

Observations

- The relative share of full RE fixed points (process perspective, Figure 5.3) is similar to the corresponding results from the result perspective (Figure 5.2) for `QuadraticGlobalRE`, `LinearGlobalRE`, and `QuadraticLocalRE` except for `LinearLocalRE`.
- For `LinearLocalRE`, the relative share of full RE fixed points is significantly higher when considering the fixed points from all branches (process perspective) rather than the set of fixed points (result perspective). This means that a relatively higher share of branches leads to full RE fixed points than to non-full-RE fixed points.
- The relative share of full RE fixed points for `LinearLocalRE` (Figure 5.3) even exceeds the relative share of full RE global optima for linear model variants (Figure 5.1).
- The number of fixed points in the process perspective (Table 5.3) is only slightly higher than the number in the result perspective (Table 5.2) for `QuadraticGlobalRE` and `LinearGlobalRE`. In contrast, the number of fixed points from all branches is substantially higher than the number of fixed points from the result perspective for `QuadraticLocalRE`, and even more so for `LinearLocalRE`.

5.2.2 Results Grouped by Sentence Pool Size

Observations

- The relative share of full RE states among global optima decreases with increasing sentence-pool size for all model variants (Figure 5.4).
- The relative share of full RE states among the set of fixed points (result perspective) decreases with increasing sentence-pool size for all model variants (Figure 5.5).
- The relative share of full RE states among the fixed points from all branches (process perspective) decreases with increasing sentence-pool size for the model variants`QuadraticLocalRE`, `QuadraticGlobalRE` and `LinearGlobalRE` (Figure 5.6).
- The relative share of full RE states among fixed points from all branches (process perspective) is roughly constant with respect to sentence pool sizes for `LinearLocalRE` (Figure 5.6).

5.2.3 Results Grouped by Configuration of Weights

Observations

- Linear model variants exhibit a "tipping line" (see Appendix A). For $\alpha_A > \alpha_F$, the relative share of full RE global optima is 1.0, i.e., all global optima are full RE states.
- Quadratic model variants have a smooth transition between low and high relative shares and have a "hotspot" for very high values of α_A. This result is made plausible by the fact that full RE states require a maximal value for the measure of account (i.e., $A(\mathcal{C}, \mathcal{T}) = 1$). High values for α_A benefit the fulfilment of this requirement.

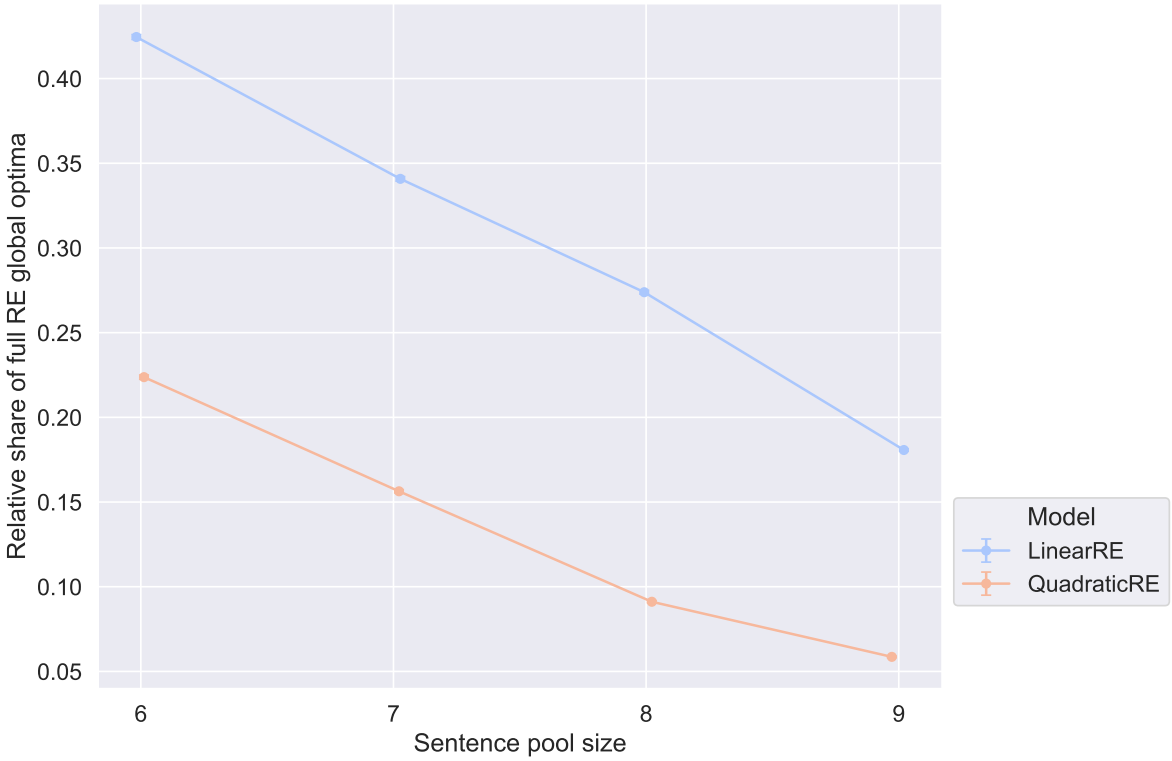


Figure 5.4: Relative share of full RE states among global optima grouped by model variant and sentence pool size

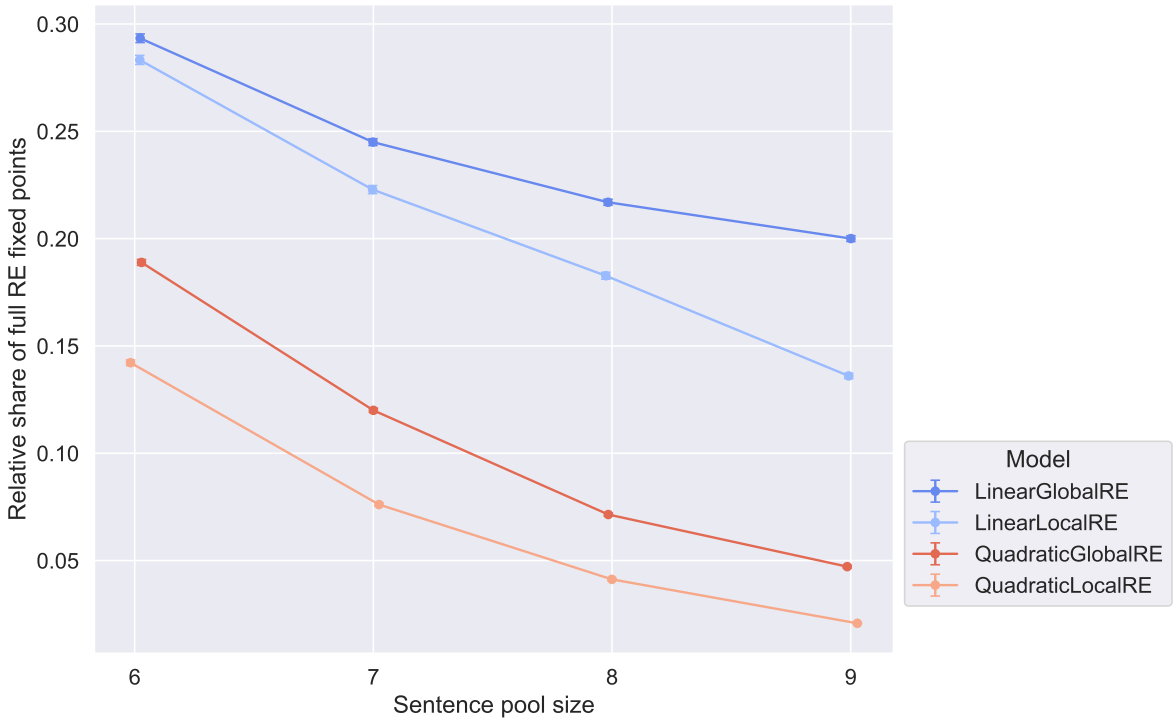


Figure 5.5: Relative share of full RE states among fixed points (result perspective) grouped by model variant and sentence pool size

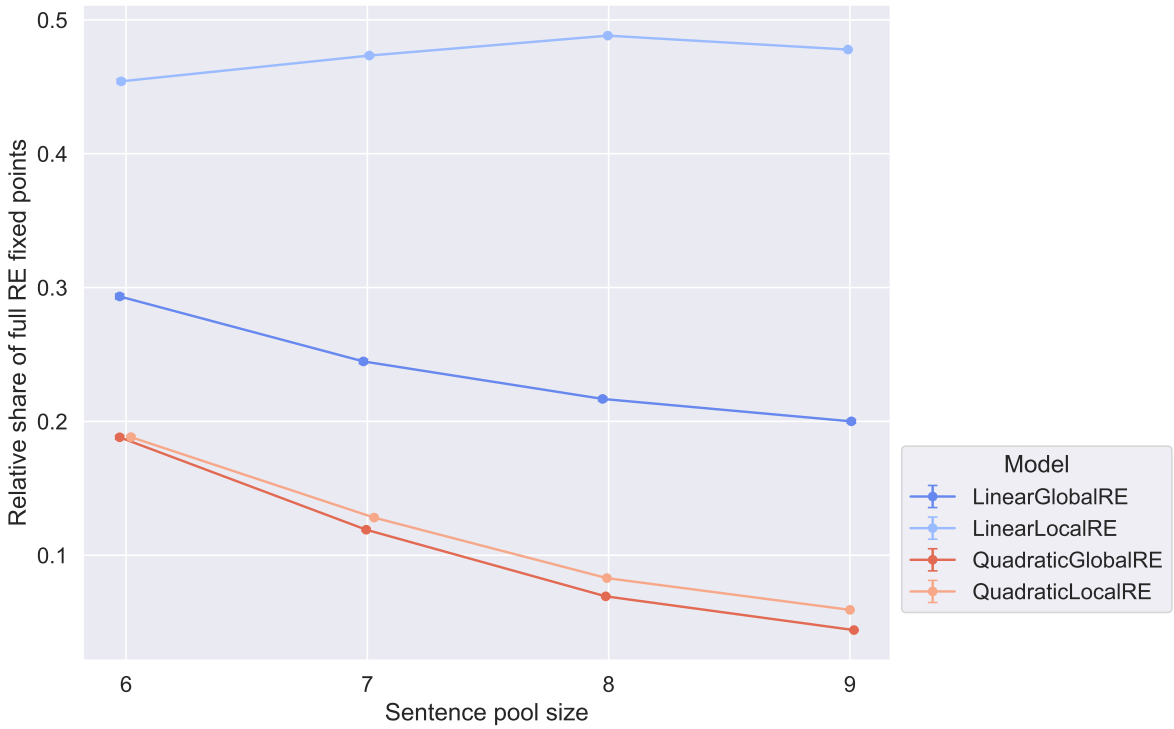


Figure 5.6: Relative share of full RE states among fixed points (process perspective) grouped by model variant and sentence pool size

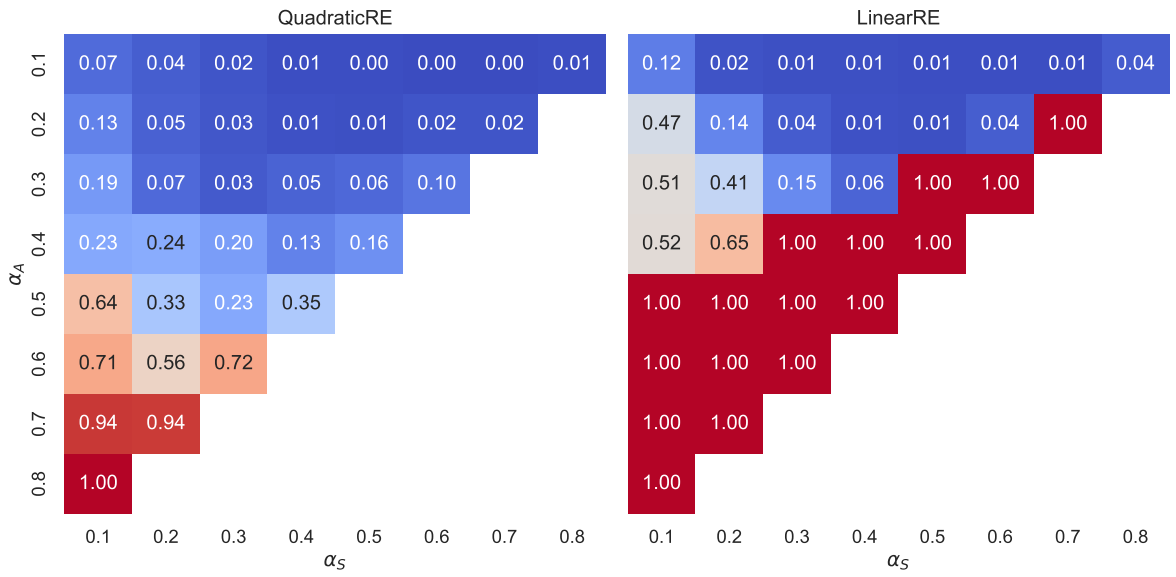


Figure 5.7: Relative share of full RE states among global optima grouped by model variant and configuration of weights

Observations

- Linear model variants do not exhibit the tipping line for fixed points (Figure 5.8 and Figure 5.9)
- Linear model variants have high relative shares for low faithfulness, moderate account and high (but non-extreme) weights for systematicity.
- There are only small differences between the relative share of full RE states among sets of fixed points (result perspective, Figure 5.8) and fixed points from all branches (process perspective, Figure 5.9).
- `QuadraticGlobalRE` exhibits its highest relative shares of full RE fixed points for moderately high values for α_A and very low values for α_S.

5.3 Conclusion

Overall, the relative share of full RE states among global optima and fixed points is not overwhelming. However, heatmaps reveal combinations of weights for `QuadraticGlobalRE`, `LinearGlobalRE` and `LinearLocalRE`, where the relative share of full RE states among the outputs is acceptable. For `QuadraticLocalRE`, this holds at least for global optima. However, this is not a strong reason to reject `QuadraticLocalRE`. Depending on the particular goals of an RE inquiry, a low relative share of full RE states can be seen as a strength of a model,

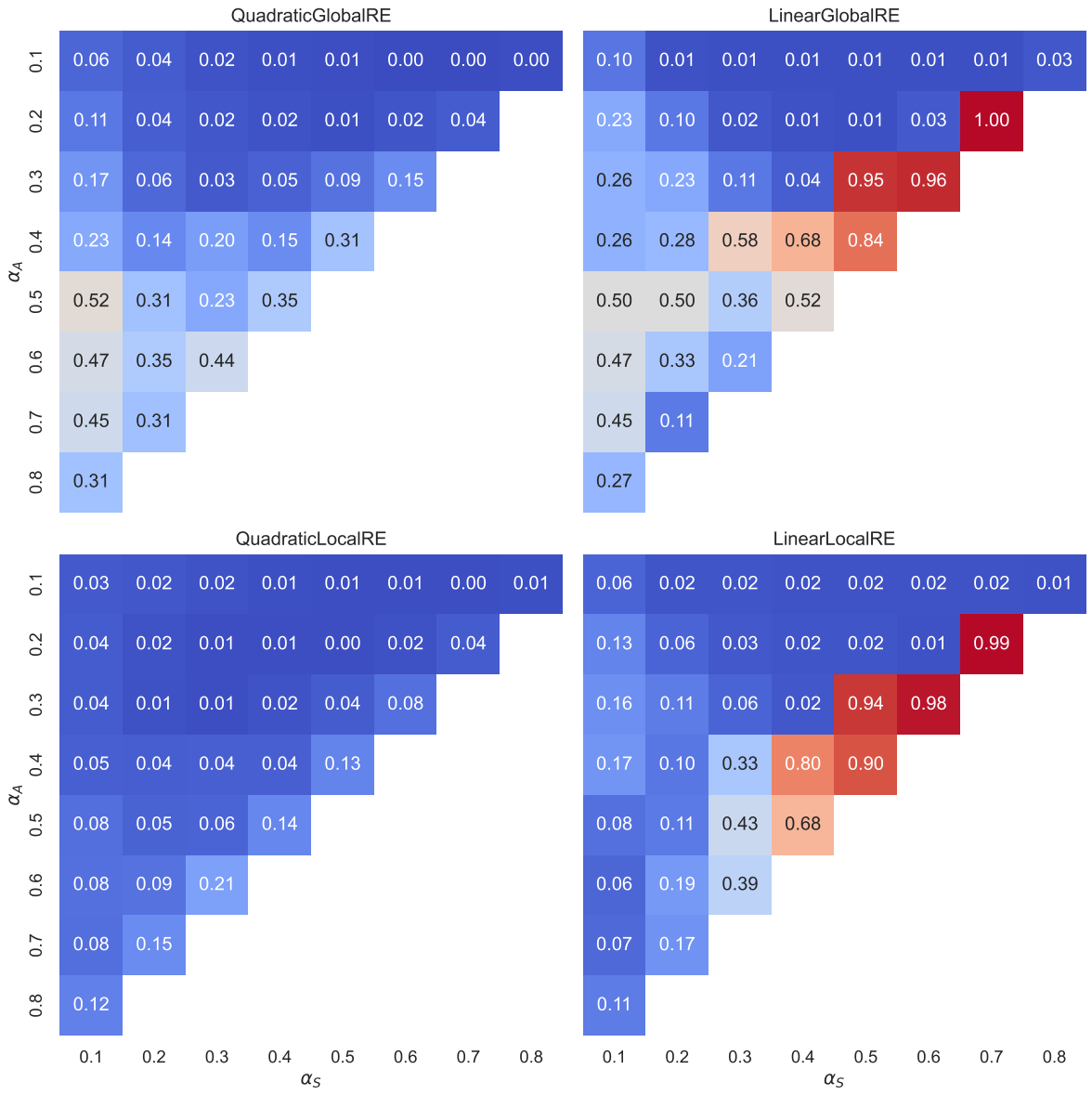


Figure 5.8: Relative share of full RE states among unique fixed points grouped by model variant and configuration of weights.

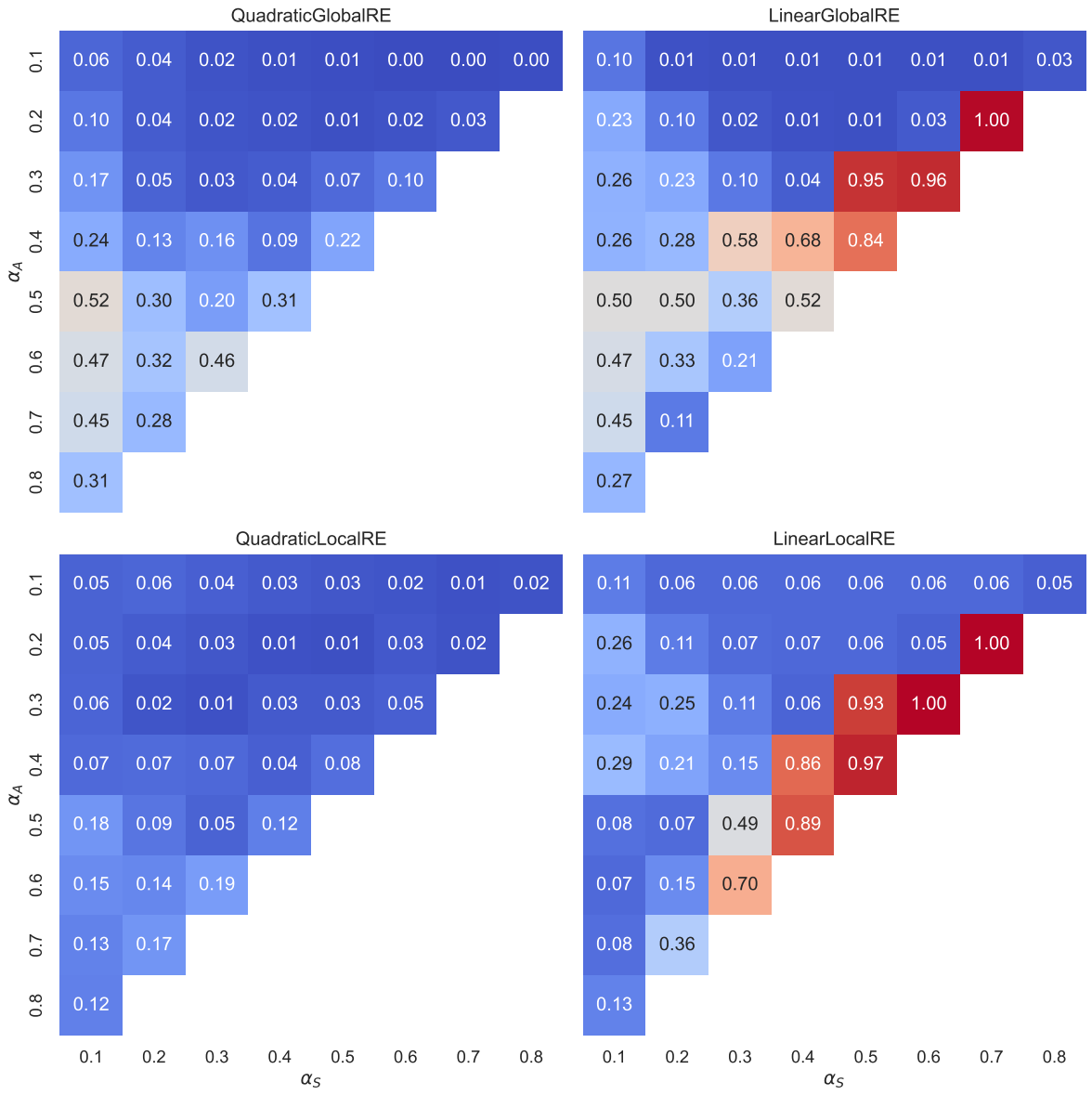


Figure 5.9: Relative share of full RE states among fixed points from all branches grouped by model variant and configuration of weights.

as it may not be desirable to render everything into a full RE state, or states satisfying less demanding requirements may be acceptable.

Concerning the influence of the sentence pool size, there is a negative trend for the relative shares of full RE states among global optima and fixed points (result perspective). Only the relative share of full RE fixed points (process perspective) of the `LinearLocalRE` model is not affected by the sentence pool size. At this point, we cannot offer an explanation for this behaviour, which calls for further analysis.

6 Consistency

6.1 Background

Consistency is commonly seen as a necessary condition of coherence. Achieving consistency in RE is, therefore, of utmost importance. In contrast to the desiderata of faithfulness, systematicity and account (see Section 2.1), the desideratum of consistency is not hard-wired into the model. Although the agent is not allowed to choose commitments with flat contradictions (i.e., commitment sets of the form $\{s_i, \ldots, \neg s_i\}$), they can choose dialectically inconsistent commitments (i.e., commitments that are inconsistent with respect to the inferential relationships encoded in the dialectical structure τ). Or, more formally, a dialectically inconsistent set of commitments may maximize the achievement during the step of adjusting commitments. Accordingly, the process might end at a fixed point with dialectically inconsistent commitments. The question is, therefore, whether the explicitly modelled desiderata and the specification of the process are sufficiently conducive towards dialectical consistency.[1]

In this chapter, we analyze the *dialectical consistency* of inputs and outputs (fixed points and global optima) of RE simulations, which can be examined from three different perspectives:

1. the consistency of output commitments
2. the "consistency case" that arises from combining the consistency status of initial and output commitments
3. the consistency of the union of output commitments and theory

Concerning 2., the juxtaposition of initial and output commitments allows for four cases, which are labelled as follows:

	endpoint commitments consistent	endpoint commitment inconsistent
initial commitments consistent	consistency preserving (CP)	consistency eliminating (CE)
initial commitments inconsistent	inconsistency eliminating (IE)	inconsistency preserving (IP)

[1] The main driving force for dialectical consistency is the desideratum of account. Since the choice of new theories is confined to dialectically consistent theories, account will favour commitments that are dialectically consistent.

CP Cases preserve or "transfer" consistency between initial and endpoint commitments. In IE cases, inconsistent initial commitments are revised for consistent endpoint commitments. IP cases fail to eradicate initial inconsistencies, and finally, there may be CE cases if inconsistencies are introduced to initially consistent commitments.

From the viewpoint of model consolidation, the cases are interesting and relevant in various respects. High shares of IE cases would speak in favour of the model's revisionary power and signify progress towards establishing coherence by RE. Frequent IP cases, in turn, would speak against the model's revisionary power with respect to inconsistent initial commitments. Moreover, this could fuel the objection that RE (or the present model thereof) is overly conservative, such that "garbage in" (inconsistent initial commitments) leads to "garbage out" (inconsistent fixed point/global optimum commitments). High relative shares of CP cases are a desirable feature. Finally, frequent CE cases would be a truly worrisome result, as they would indicate that the model leads to a worsening in terms of consistency.

6.2 Results

> **Note**
>
> The results of this chapter can be reproduced with the following Jupyter notebook: https://github.com/re-models/re-technical-report/blob/main/notebooks/data_analysis_chapter_commitment-consistency.ipynb.

6.2.1 Consistent Outputs

6.2.1.1 Overall Results

Model	Relative share of global optima with consistent commitments	Number of global optima with consistent commitments	Number of global optima
QuadraticGlobalRE	0.741	529359	714584
LinearGlobalRE	0.771	540556	700830
QuadraticLocalRE	0.741	525490	709289
LinearLocalRE	0.769	554525	721096

Table 6.2: Relative share of consistent commitments among global optima

Model	Relative share of fixed points with consistent commitments	Number of fixed points with consistent commitments	Number of fixed points
QuadraticGlobalRE	0.728	333436	458147
LinearGlobalRE	0.726	227000	312783
QuadraticLocalRE	0.688	404941	588236
LinearLocalRE	0.82	187163	228122

Table 6.3: Relative share of consistent commitments among fixed points (result perspective)

Model	Relative share of fixed points with consistent commitments	Number of fixed points with consistent commitments	Number of fixed points
QuadraticGlobalRE	0.708	374476	528616
LinearGlobalRE	0.726	227097	313002
QuadraticLocalRE	0.735	1463131	1991852
LinearLocalRE	0.952	1240692	1303077

Table 6.4: Relative share of consistent commitments among fixed points (process perspective)

Observations: Consistent Outputs

- Overall, the relative share of consistent output commitments is high for all model variants and output types, roughly ranging from 0.69 to 0.95
- The overall relative share of consistent global optima commitments is slightly boosted for linear model variants compared to their quadratic counterparts in Table 6.2.
- The relative shares of consistent commitments among fixed points (result perspective: Table 6.3, and process perspective: Table 6.4) is slightly lower than the corresponding results for global optima in Table 6.2 for `QuadraticGlobalRE`, `QuadraticLocalRE`, and `LinearGlobalRE`
- `LinearLocalRE` exhibits substantially higher relative shares of consistent commitments among fixed points (result and process perspective)
- The number of fixed points reached through different branches (process perspective) in local model variants is substantially higher than for global model variants (Table 6.4)

6.2.1.2 Results Grouped by Sentence Pool Size

Observations

- The relative share of global optima with consistent commitments slighty decrease for larger sentence pool sizes (Figure 6.4).

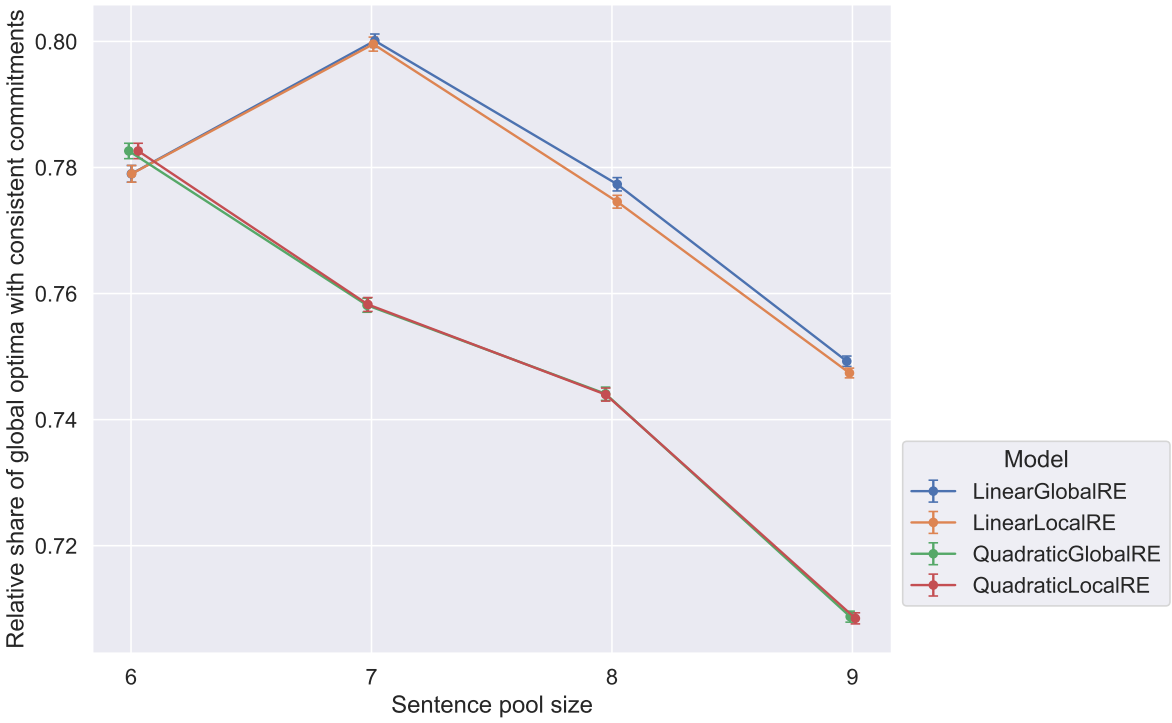


Figure 6.1: Relative share of global optima with consistent commitments grouped by model variant and sentence pool size

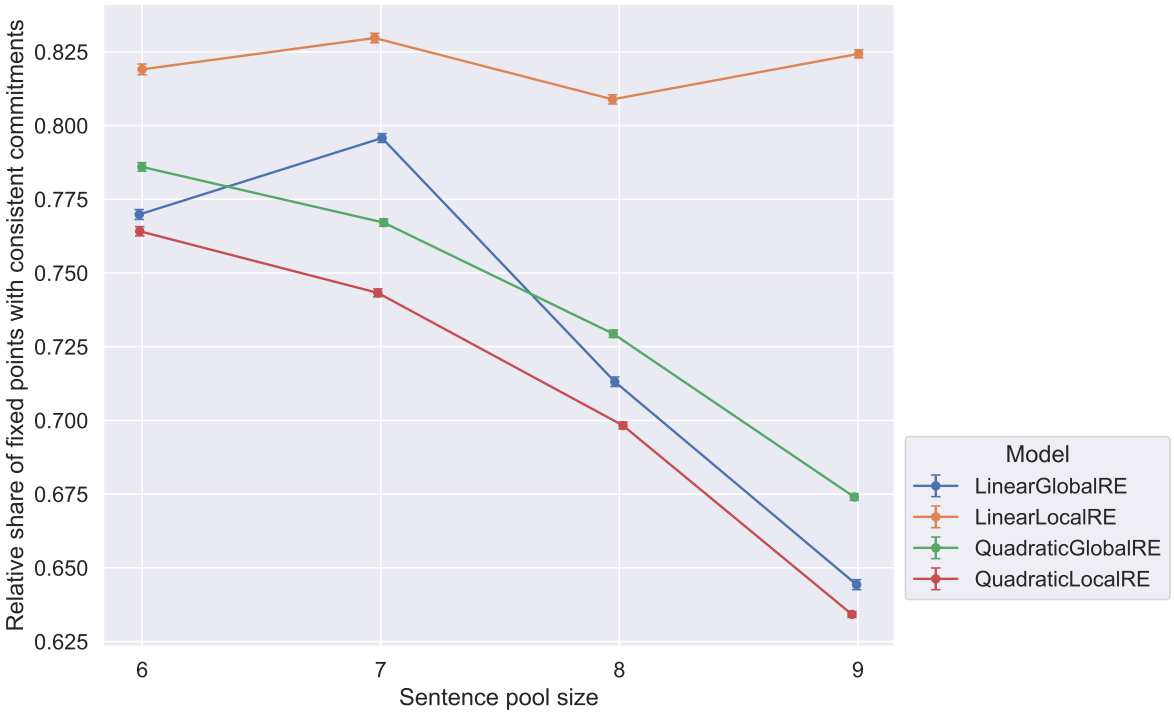


Figure 6.2: Relative share of fixed points (result perspective) with consistent commitments grouped by model variant and sentence pool size

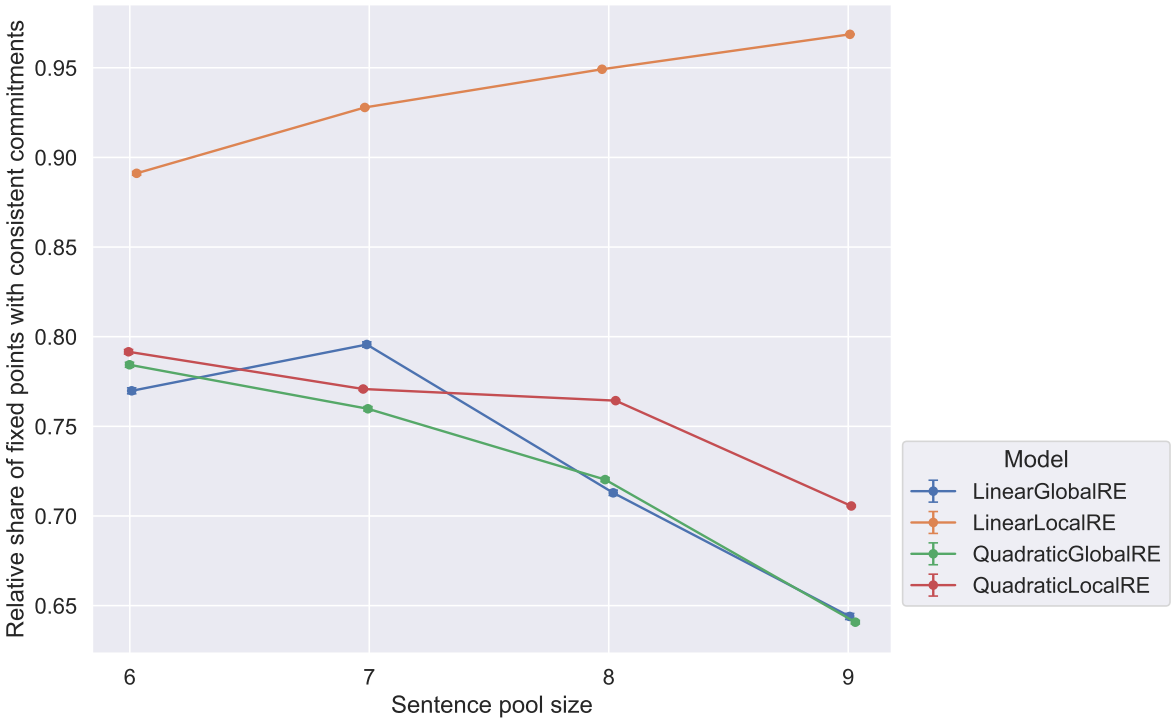


Figure 6.3: Relative share of fixed points (process perspective) with consistent commitments grouped by model variant and sentence pool size

- The closeness of results of `QuadraticGlobalRE` and `QuadraticLocalRE`, as well as `LinearGlobalRE` and `LinearLocalRE` in Figure 6.4 is due to the fact, that local variants rely on their global counterparts to determine global optima. Differences arise due to the exclusion of different erroneous runs.
- The relative share of fixed points with consistent commitments slightly decreases for larger sentence pool sizes (both perspectives in Figure 6.5 and Figure 6.6) for `QuadraticGlobalRE`, `QuadraticLocalRE`, and `LinearGlobalRE`.
- In contrast, for `LinearLocalRE`, the relative share of fixed points with consistent commitments remains roughly constant (result perspective in Figure 6.5) or sligtly increases (process perspective in Figure 6.6)

6.2.1.3 Results Grouped by Configuration of Weights

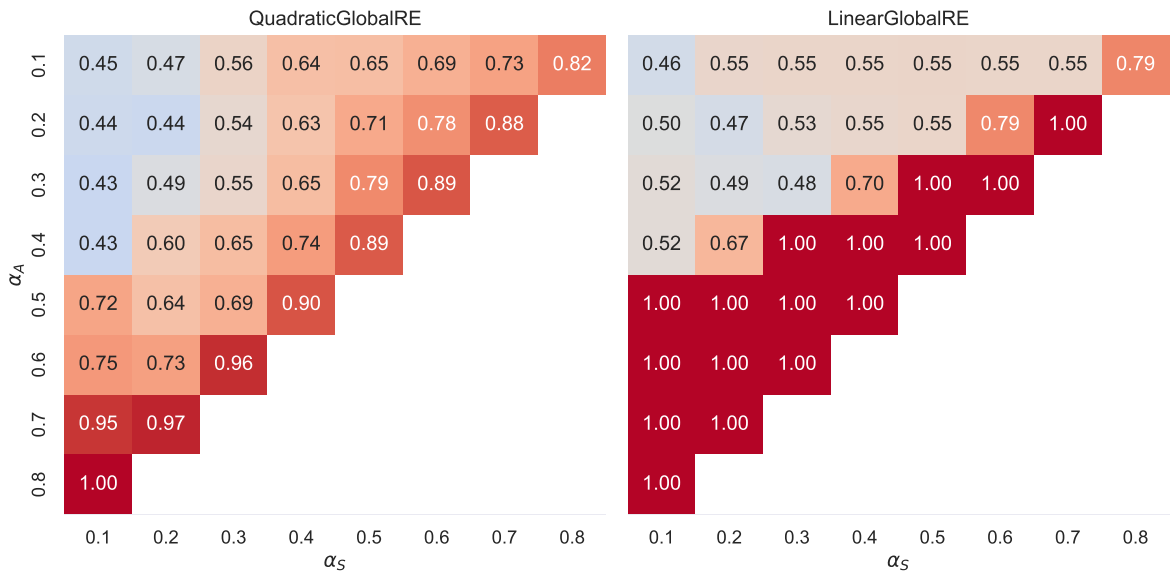


Figure 6.4: Relative share of global optima with consistent commitments grouped by model variant and configuration of weights. Note that local variants are omitted due to almost analogous results.

Observations

- Linear models exhibit a "tipping line" for the relative share of global optima and fixed points with consistent commitments. For $\alpha_A > \alpha_F$, the relative share is consistently 1.0. See Appendix A for an explanation.
- In contrast, quadratic models show a gradient of smoother transitions between relative shares, increasing with higher weights for α_A, and also to some extent with higher weights for α_A.

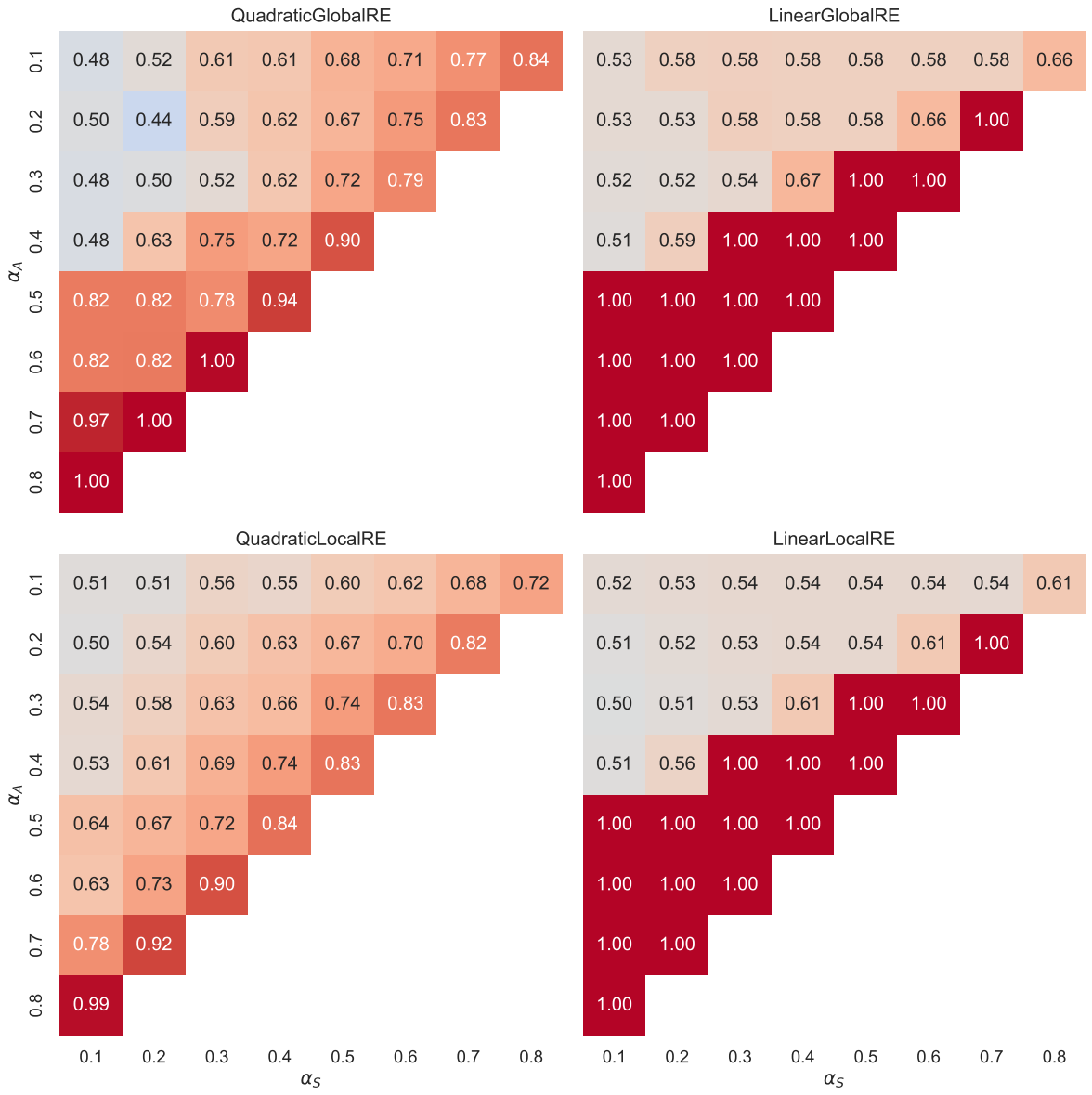


Figure 6.5: Relative share of fixed points (result perspective) with consistent commitments grouped by model variant and configuration of weights

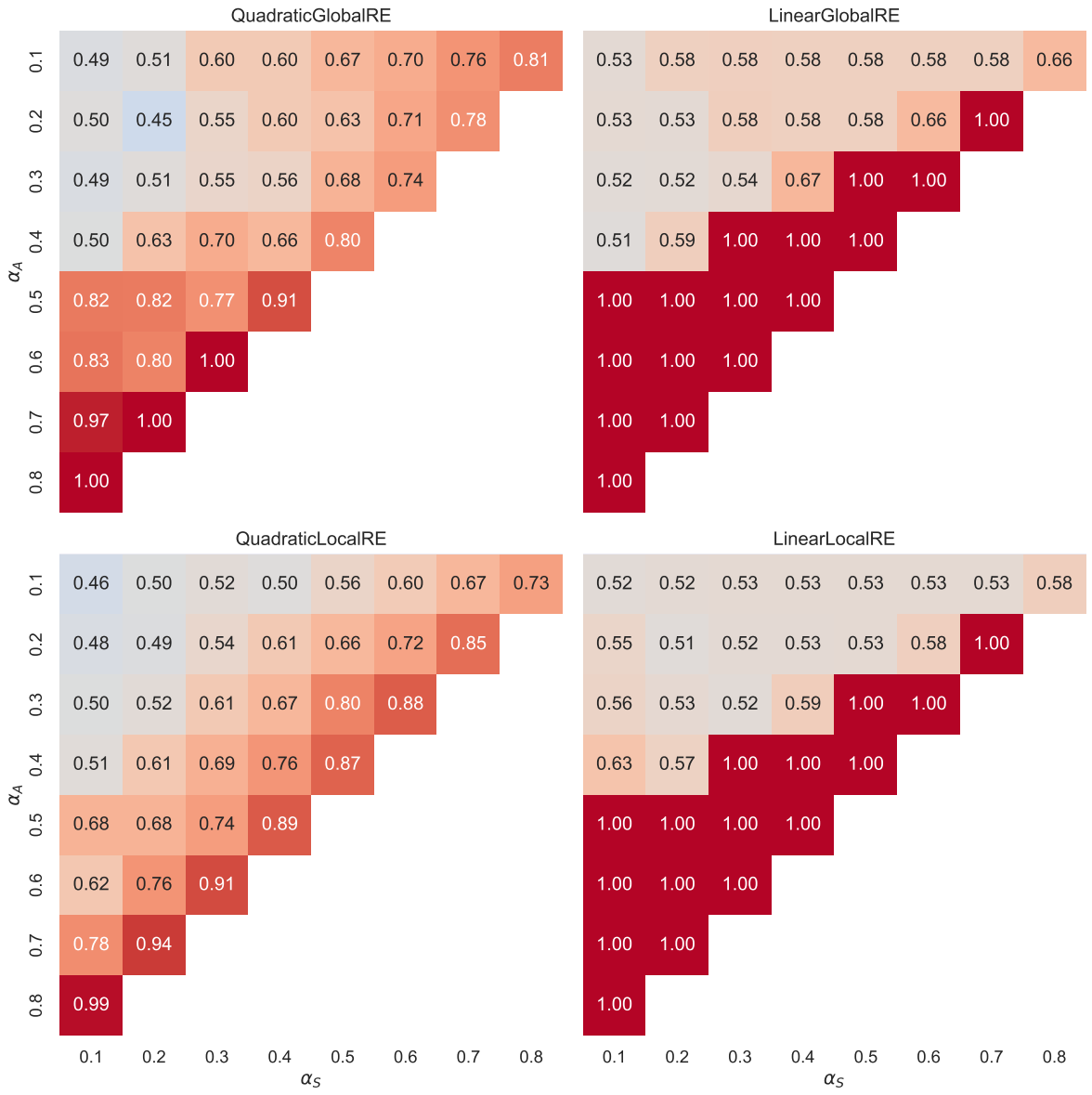


Figure 6.6: Relative share of fixed points (process perspective) with consistent commitments grouped by model variant and configuration of weights

6.2.2 Consistency Cases

The results of this section are based on a more fine-grained distinction of cases that depend on the consistency status of initial and output commitments.

Note that the relative shares of cases have been calculated for consistent and inconsistent initial commitments separately. For example, the relative share of inconsistency eliminating cases (inconsistent input, consistent output) among global optima has been calculated with respect to all global optima that result from inconsistent inital commitments.

Consequently, the relative share of inconsistency eleminating and inconsistency preserving cases add up to 1.0, and so do the relative shares of consistency preserving and consistency eliminating cases.

6.2.2.1 Overall Results

Model	Relative share of consistency eliminating cases	Relative share of consistency preserving cases	Number of global optima from consistent initial commitments	Relative share of inconsistency preserving cases	Relative share of inconsistency eliminating cases	Number of global optima from inconsistent initial commitments
QGRE	0.053	0.947	386131	0.501	0.499	328453
LGRE	0.024	0.976	366296	0.453	0.547	334534
QLRE	0.053	0.947	384850	0.504	0.496	324439
LLRE	0.023	0.977	372362	0.453	0.547	348734

Table 6.5: Relative share of consistency cases among global optima

Model	Relative share of consistency eliminating cases	Relative share of consistency preserving cases	Number of fixed points from consistent initial commitments	Relative share of inconsistency preserving cases	Relative share of inconsistency eliminating cases	Number of fixed points from inconsistent initial commitments
QGRE	0.041	0.959	246823	0.543	0.457	211324
LGRE	0.016	0.984	168946	0.577	0.423	143837
QLRE	0.045	0.955	278450	0.552	0.448	309786
LLRE	0.014	0.986	119476	0.361	0.639	108646

Table 6.6: Relative share of consistency cases among fixed points (result perspective)

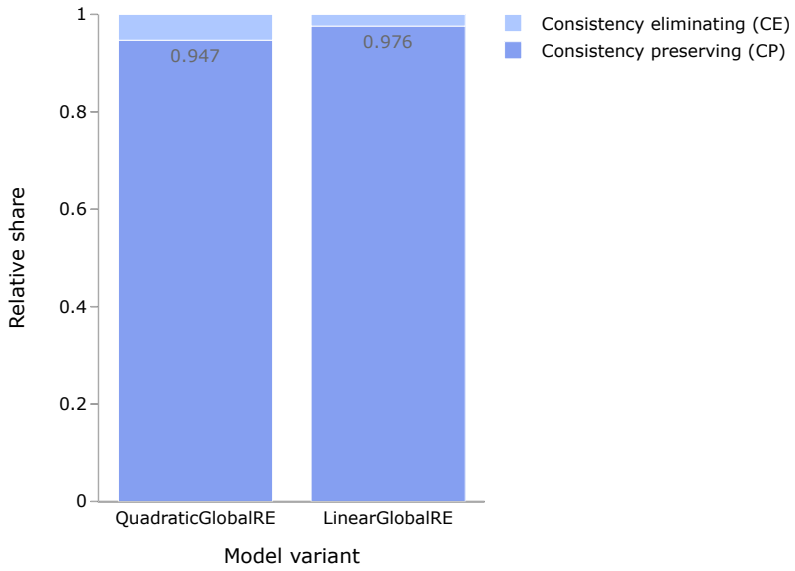


Figure 6.7: Relative share of consistency cases among global optima resulting from consistent initial commitments

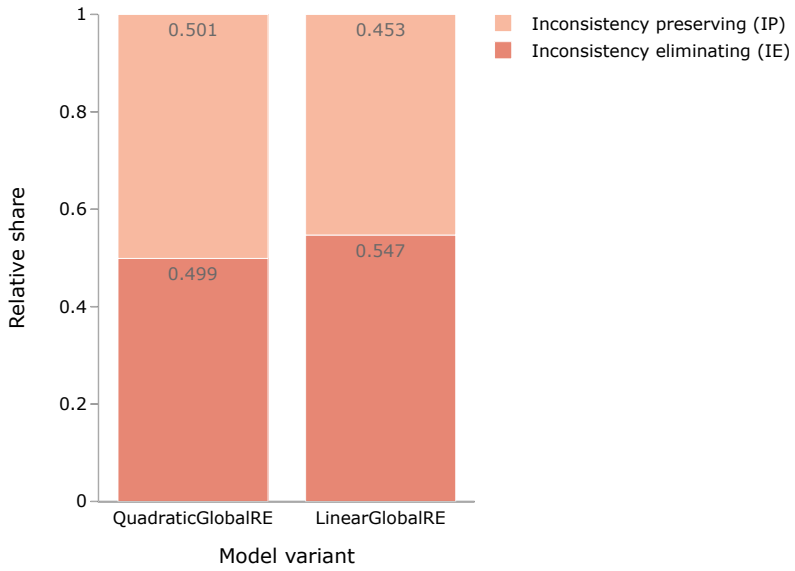


Figure 6.8: Relative share of consistency cases among global optima resulting from inconsistent initial commitments

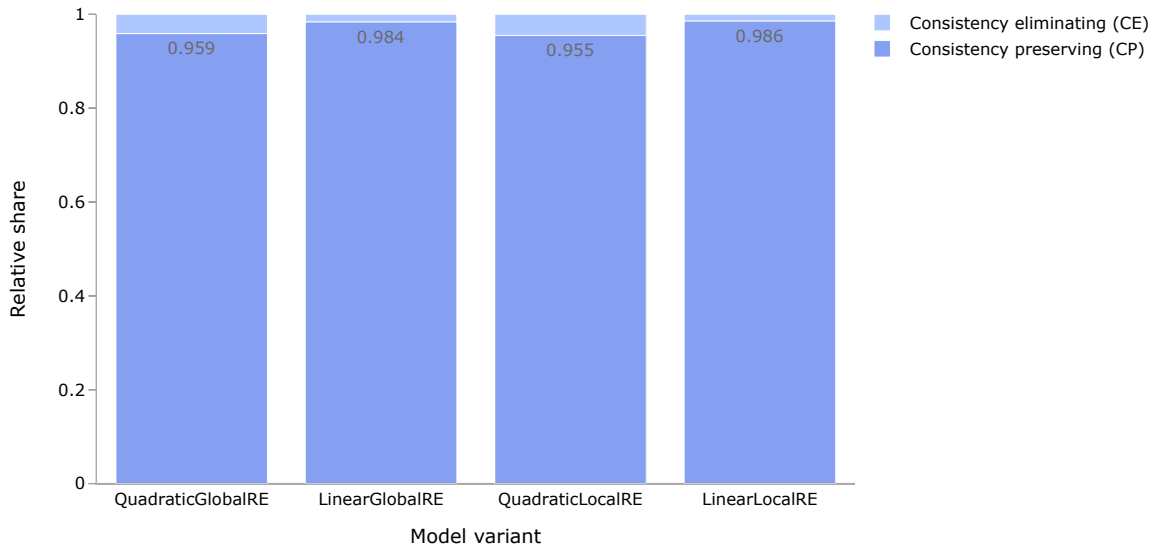


Figure 6.9: Relative share of consistency cases among fixed points (result perspective) from consistent initial commitments

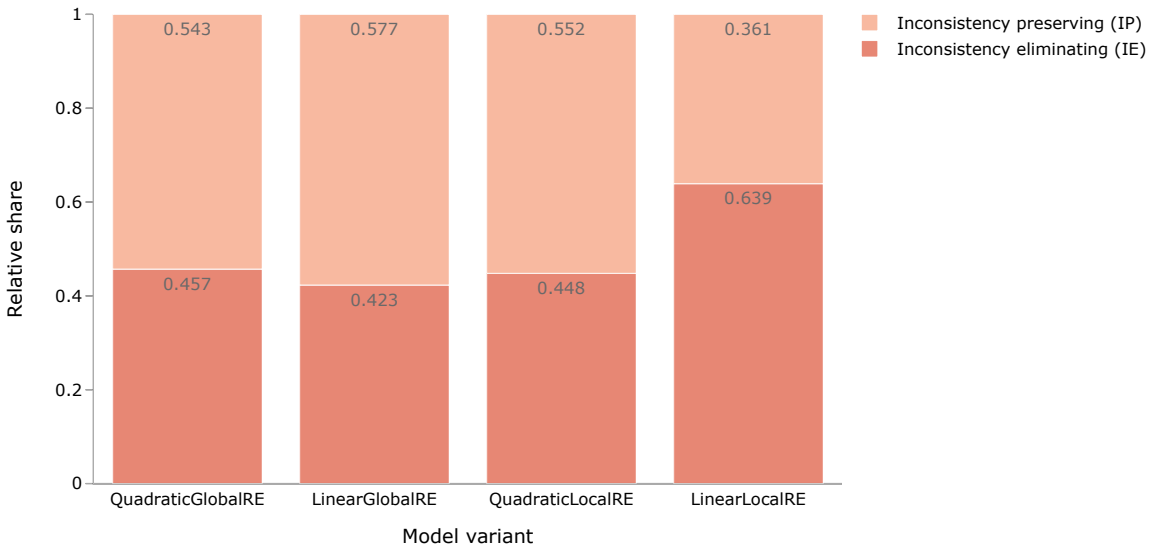


Figure 6.10: Relative share of consistency cases among fixed points (result perspective) from inconsistent initial commitments

Model	Relative share of consistency eliminating cases	Relative share of consistency preserving cases	Number of fixed points from consistent initial commitments	Relative share of inconsistency preserving cases	Relative share of inconsistency eliminating cases	Number of fixed points from inconsistent initial commitments
QGRE	0.043	0.957	264780	0.541	0.459	263836
LGRE	0.016	0.984	169026	0.578	0.422	143976
QLRE	0.057	0.943	916286	0.443	0.557	1075566
LLRE	0.006	0.994	615748	0.085	0.915	687329

Table 6.7: Relative share of consistency cases among fixed points (process perspective)

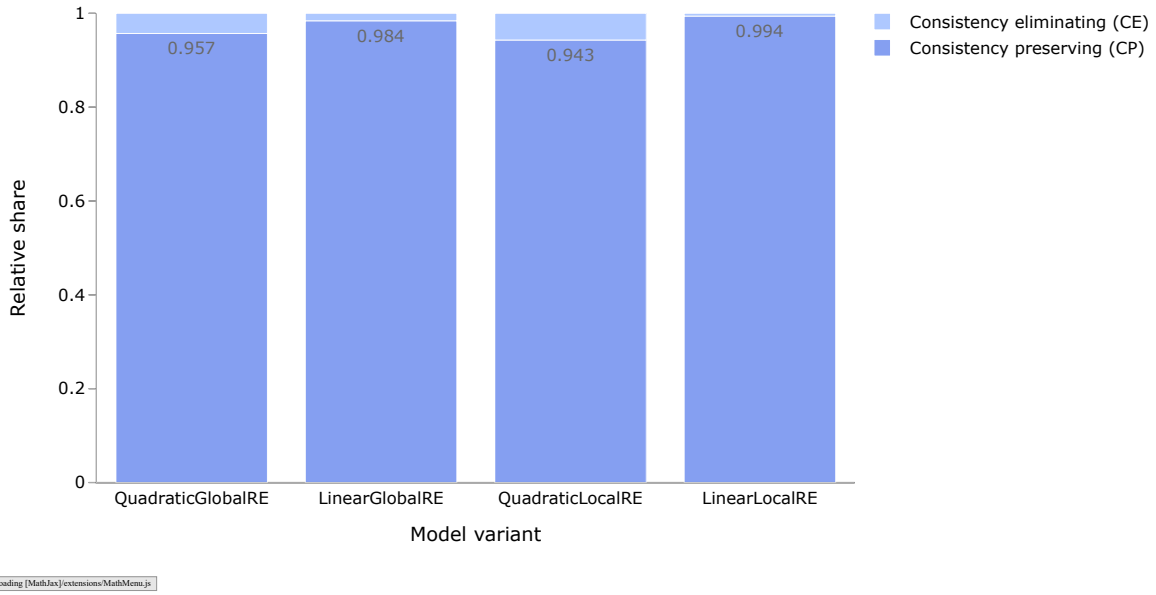


Figure 6.11: Relative share of consistency cases among fixed points (process perspective) from consistent initial commitments

Observations: Consistency Cases

- The relative share of consistency-preserving cases is high for all model variants and output types (Figure 6.7,Figure 6.9, and Figure 6.11). Consistency-eliminating cases occur very rarely.

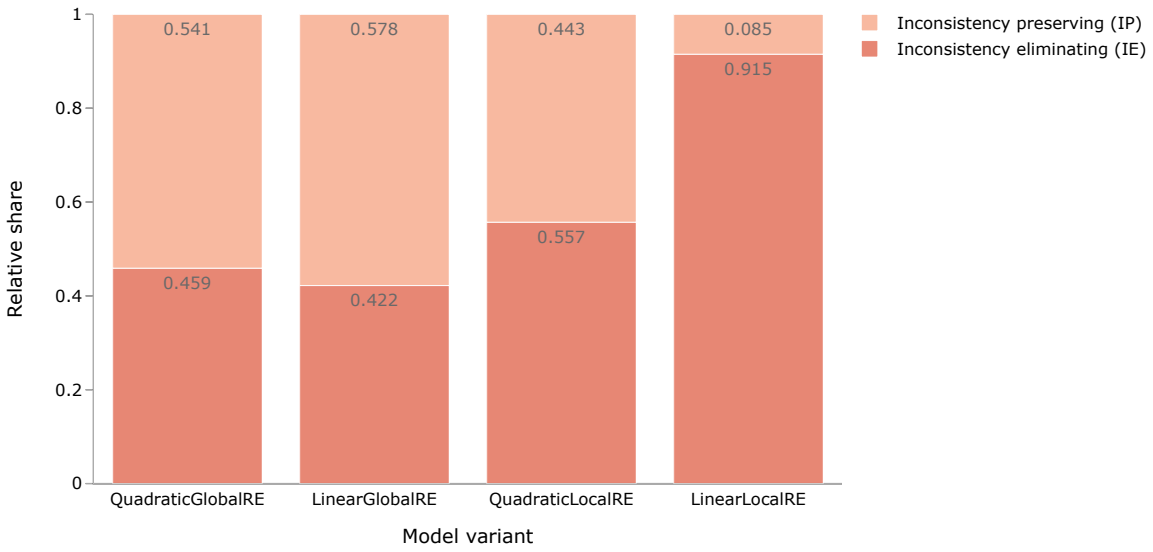


Figure 6.12: Relative share of consistency cases among fixed points (process perspective) from inconsistent initial commitments

- The relative share of inconsistency preserving cases slightly exceed the inconsistency eliminating cases for global optima and fixed points of `QuadraticGlobalRE`, `QuadraticLocalRE`, as well as `LinearGlobalRE` (Figure 6.8, Figure 6.10, and Figure 6.12).
- The result perspective makes clear that the linear local model variant reaches inconsistent output commitments from both consistent and inconsistent initial commitments (Figure 6.9 and Figure 6.10), but the process perspective reveals that only very few branches result in these inconsistent output commitments (Figure 6.11 and Figure 6.12).

6.2.2.2 Results Grouped by Sentence Pool Size

Inconsistency Eliminating Cases

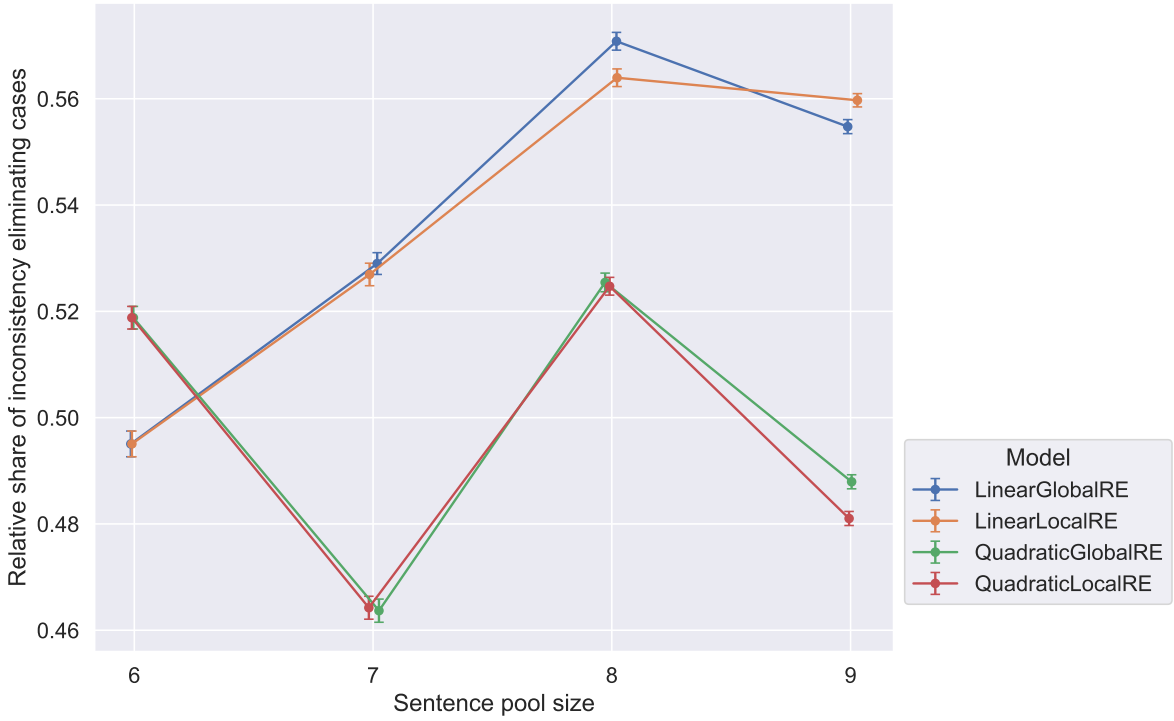


Figure 6.13: Relative share of inconsistency eliminating cases among global optima grouped by model variant and sentence pool size

Consistency Preserving Cases

Observations

- `LinearLocalRE` is the only model that tends to perform better with increasing sentence pool sizes with respect to all output types and conistency cases.

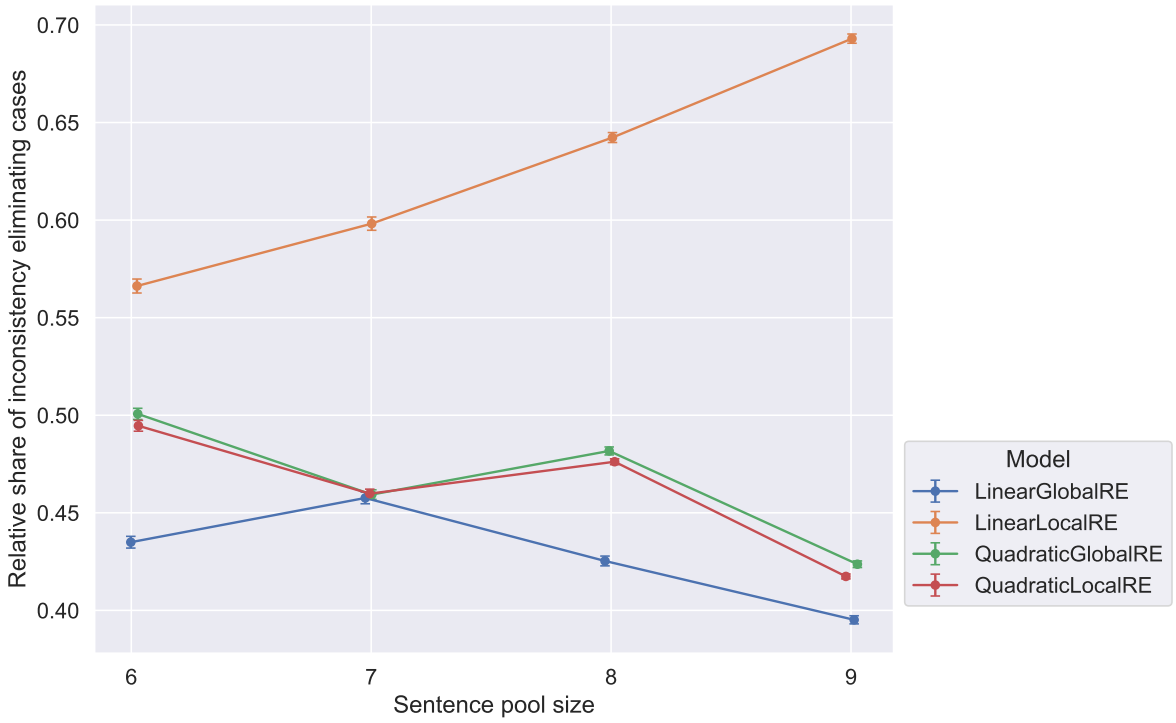


Figure 6.14: Relative share of inconsistency eliminating cases among fixed points (result perspective) grouped by model variant and sentence pool size

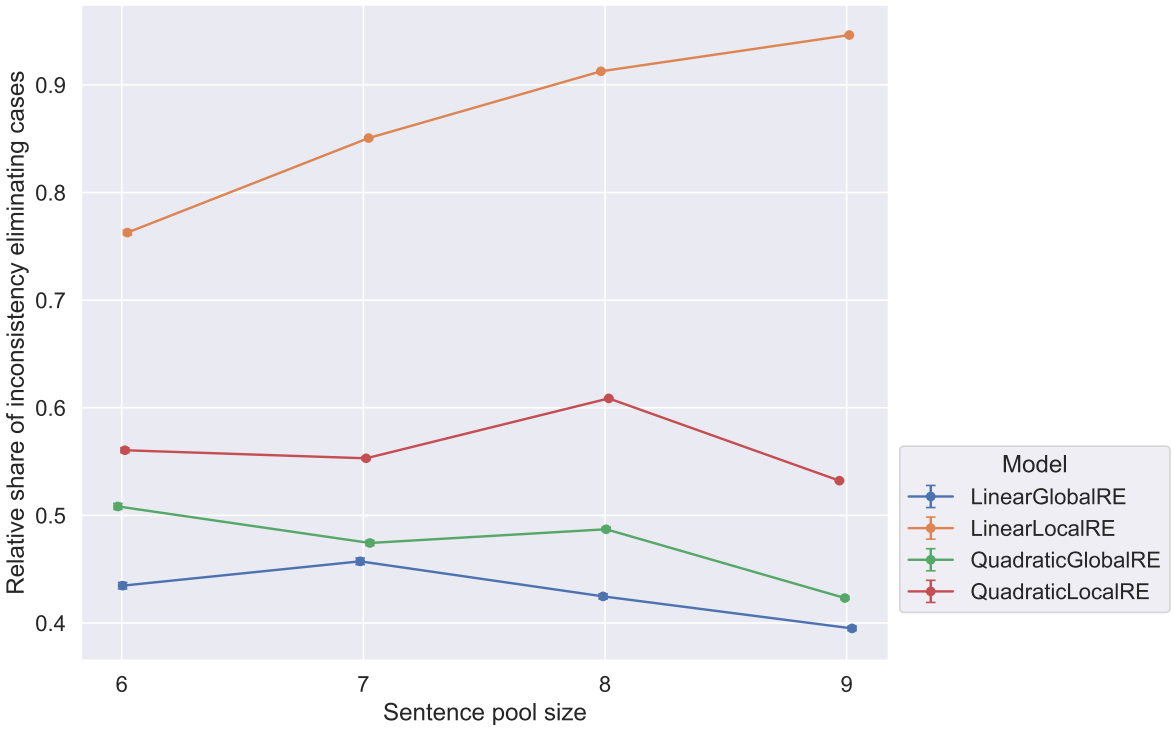


Figure 6.15: Relative share of inconsistency eliminating cases among fixed points (process perspective) grouped by model variant and sentence pool size

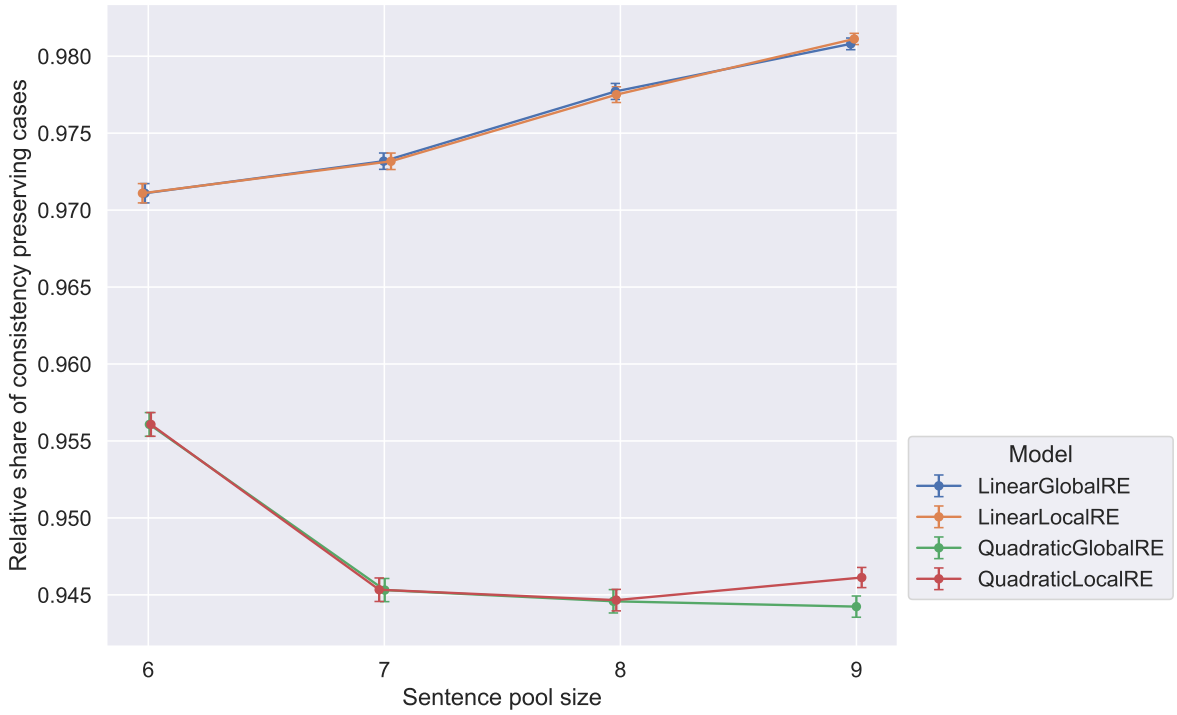


Figure 6.16: Relative share of consistency preserving cases among global optima grouped by model variant and sentence pool size

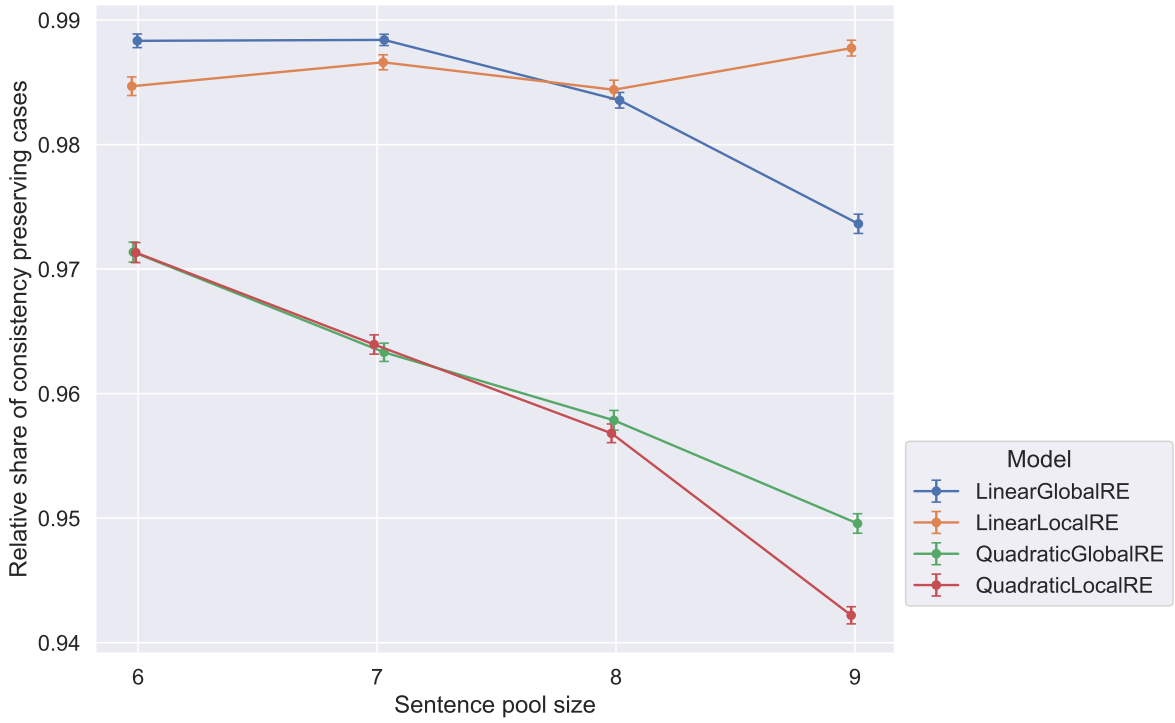


Figure 6.17: Relative share of consistency preserving cases among fixed points (result perspective) grouped by model variant and sentence pool size

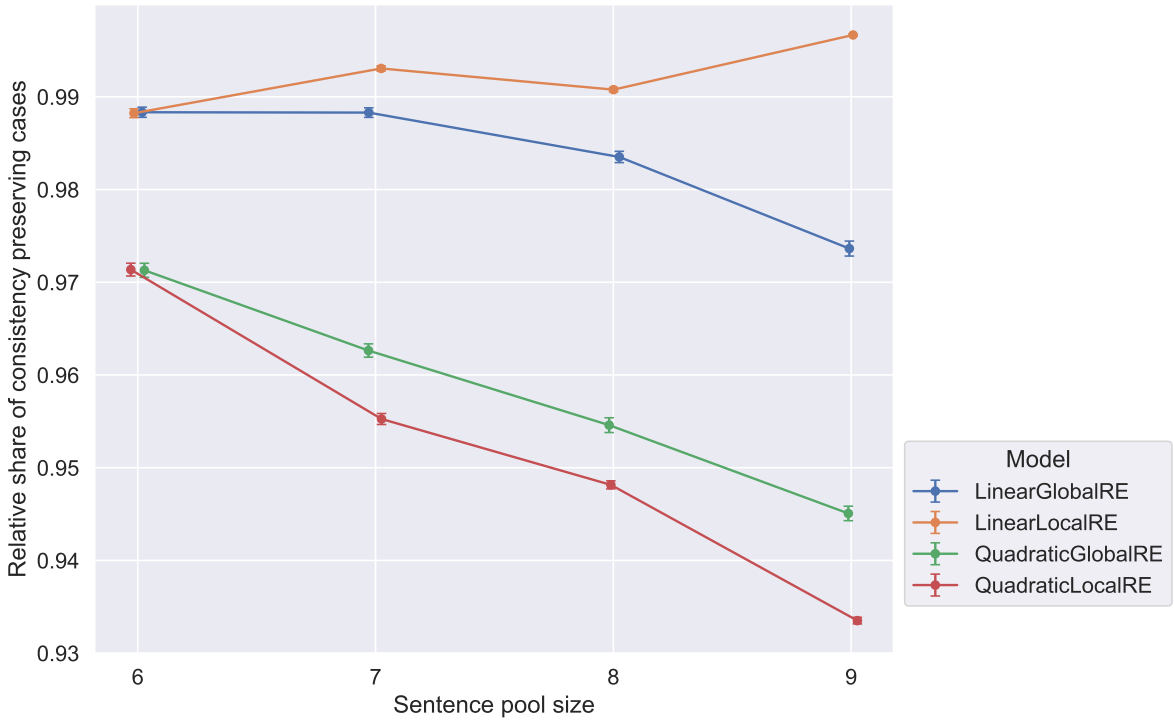


Figure 6.18: Relative share of consistency preserving cases among fixed points (process perspective) grouped by model variant and sentence pool size

6.2.2.3 Results Grouped by Configuration of Weights

Due to the fact, that inconsistency eliminating and inconsistency preserving cases, as well as consistency eliminating and consistency preserving cases are complementary, we confine the presentation of results to two cases.

Inconsistency Eliminating Cases

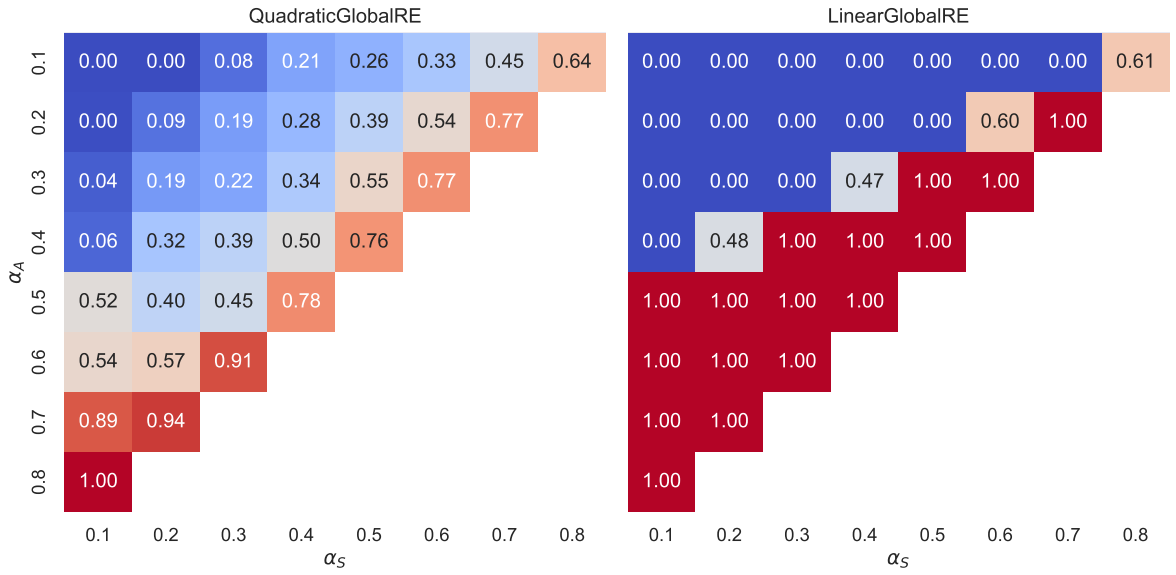


Figure 6.19: Relative share of inconsistency eliminating cases among global optima grouped by model variant and configuration of weights.

Observations: Inconsistency eliminating cases (IE)

- Linear models exhibit a "tipping line" for IE cases among both global optima and fixed points. There are no IE cases where $\alpha_A < \alpha_F$, i.e. initial inconsistencies are never removed. In turn, the relative share of IE cases for $\alpha_A > \alpha_F$ is 1.0, i.e. initial inconsistencies are always removed. See Appendix A for an explanation.
- The case with non extreme values in linear models occur where $\alpha_A = \alpha_F$.
- In contrast, quadratic models have smooth transitions. High weights for account and systematicity, resulting in low weights for faithfulness, benefit the relative share of IE cases among global optima and fixed points.
- The relative shares of IE cases among fixed points (process perspective) in local model variants (Figure 6.21) are slightly boosted in comparison to the consideration of unique fixed points (result perspectve) (Figure 6.20).

Consistency Preserving Case (CP)

Observations: Consistency Preserving Cases (CP)

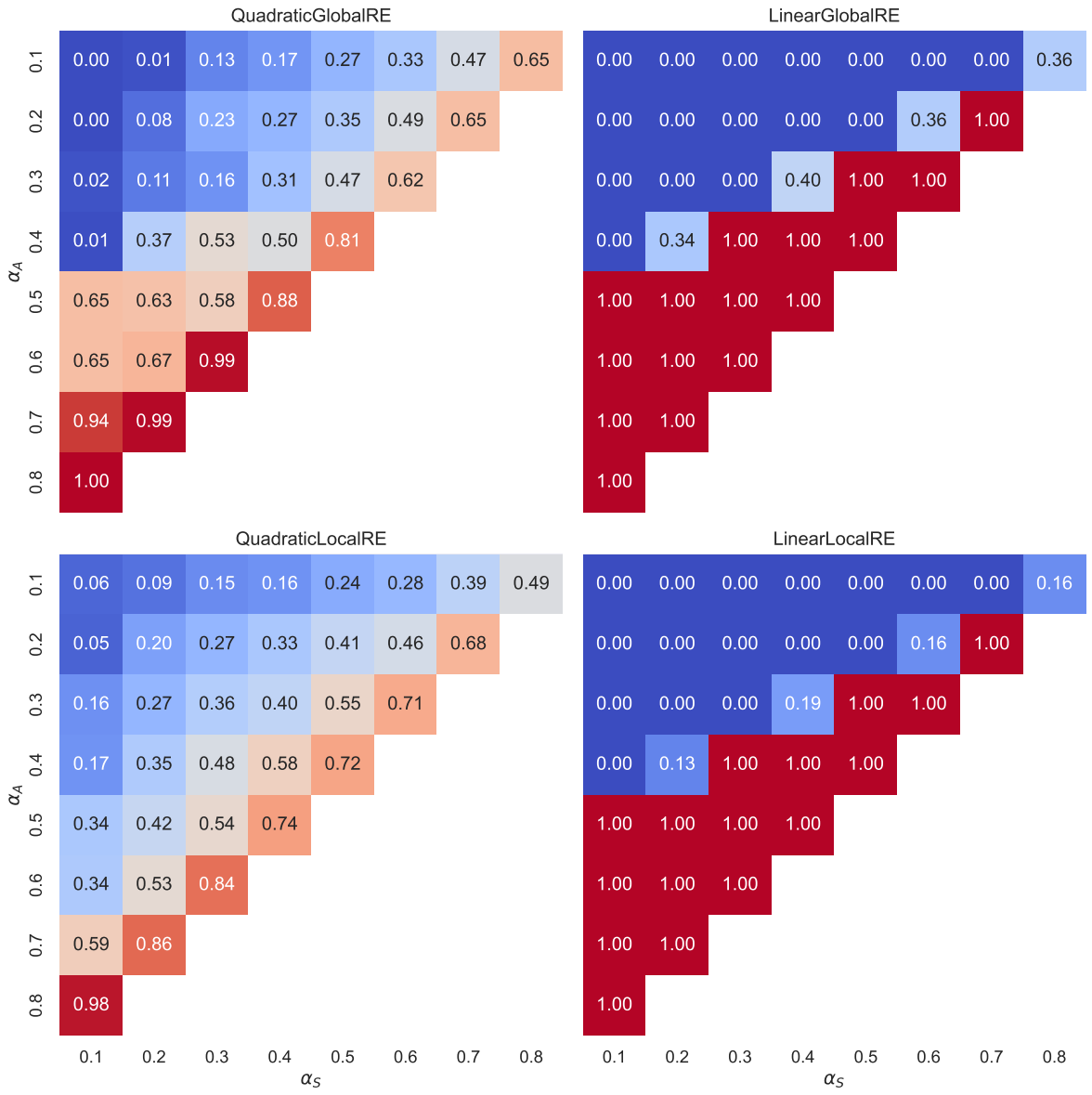


Figure 6.20: Relative share of inconsistency eliminating cases among fixed points (result perspective) grouped by model variant and configuration of weights.

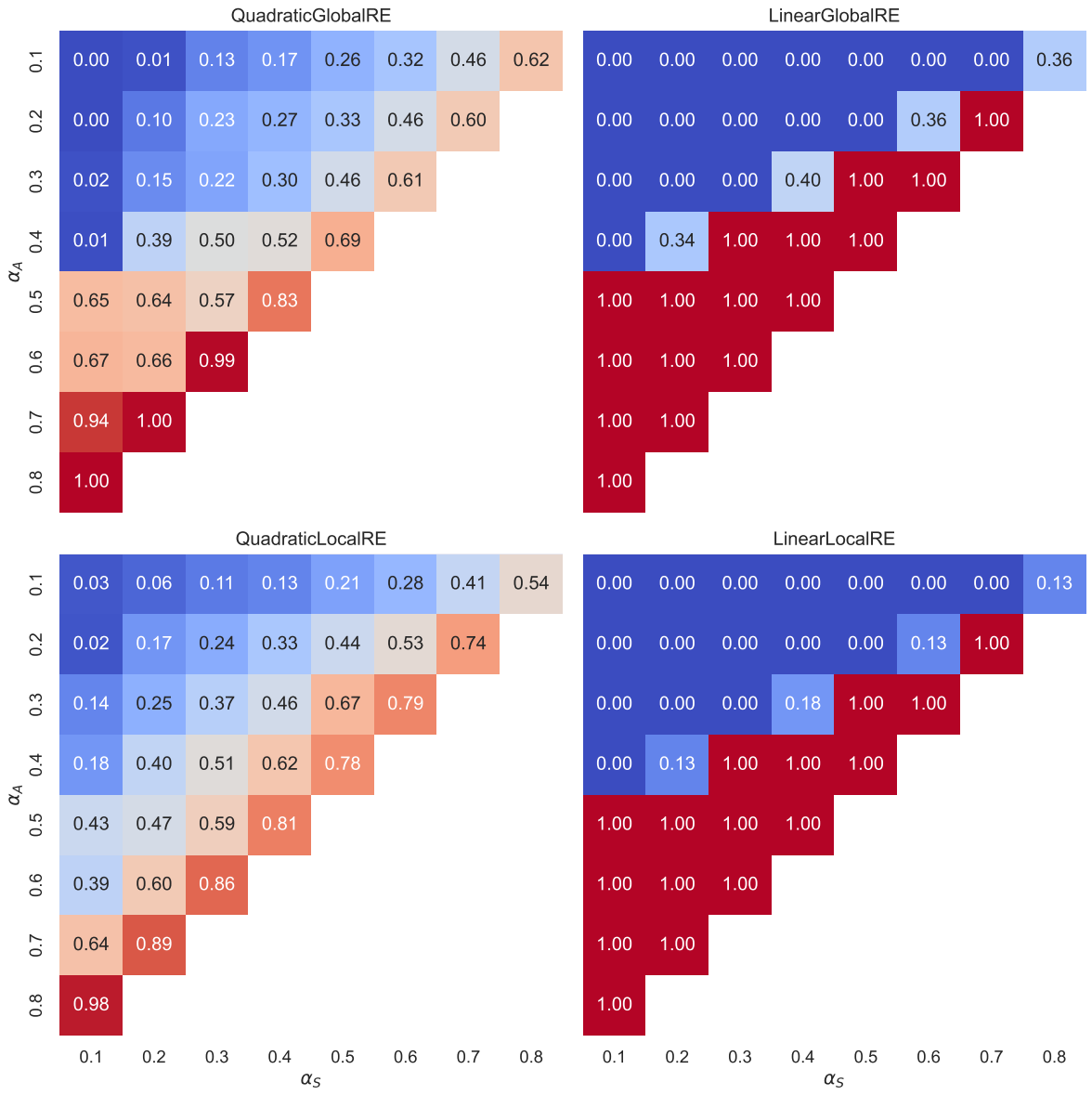


Figure 6.21: Relative share of inconsistency eliminating cases among fixed points (process perspective) grouped by model variant and configuration of weights.

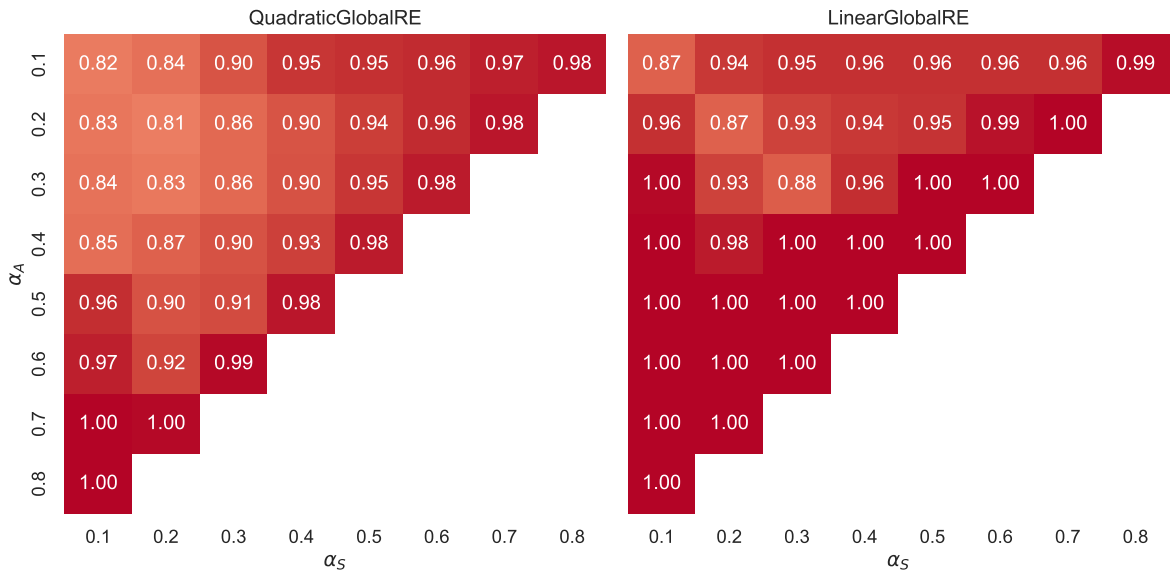


Figure 6.22: Relative share of consistency preserving cases among global optima grouped by model variant and configuration of weights.

- Overall, CP cases occur very frequently for all model variants and output types. In turn, the relative shares of CE cases ($1.0 - CP$) are very low.
- Linear models exhibit a "tipping line" for CP cases among both global optima and fixed points. For $\alpha_A > \alpha_F$, consitency is always preserved. In turn, CE cases occur only for $\alpha_A \leq \alpha_F$.
- The influence of weight configurations is moderately at best.

6.2.3 Consistent Unions

In this section, we will analyze the dialectical consistency of whole epistemic states—that is, the union of an epistemic state's commitments and theory. Since we already analyzed the consistency of fixed point commitments and global optima commitments in isolation, we will count only those inconsistencies that arise by combining commitments and theories. In other words, we will not consider inconsistencies that result from inconsistencies in the commitments.

6.2.3.1 Overall Results

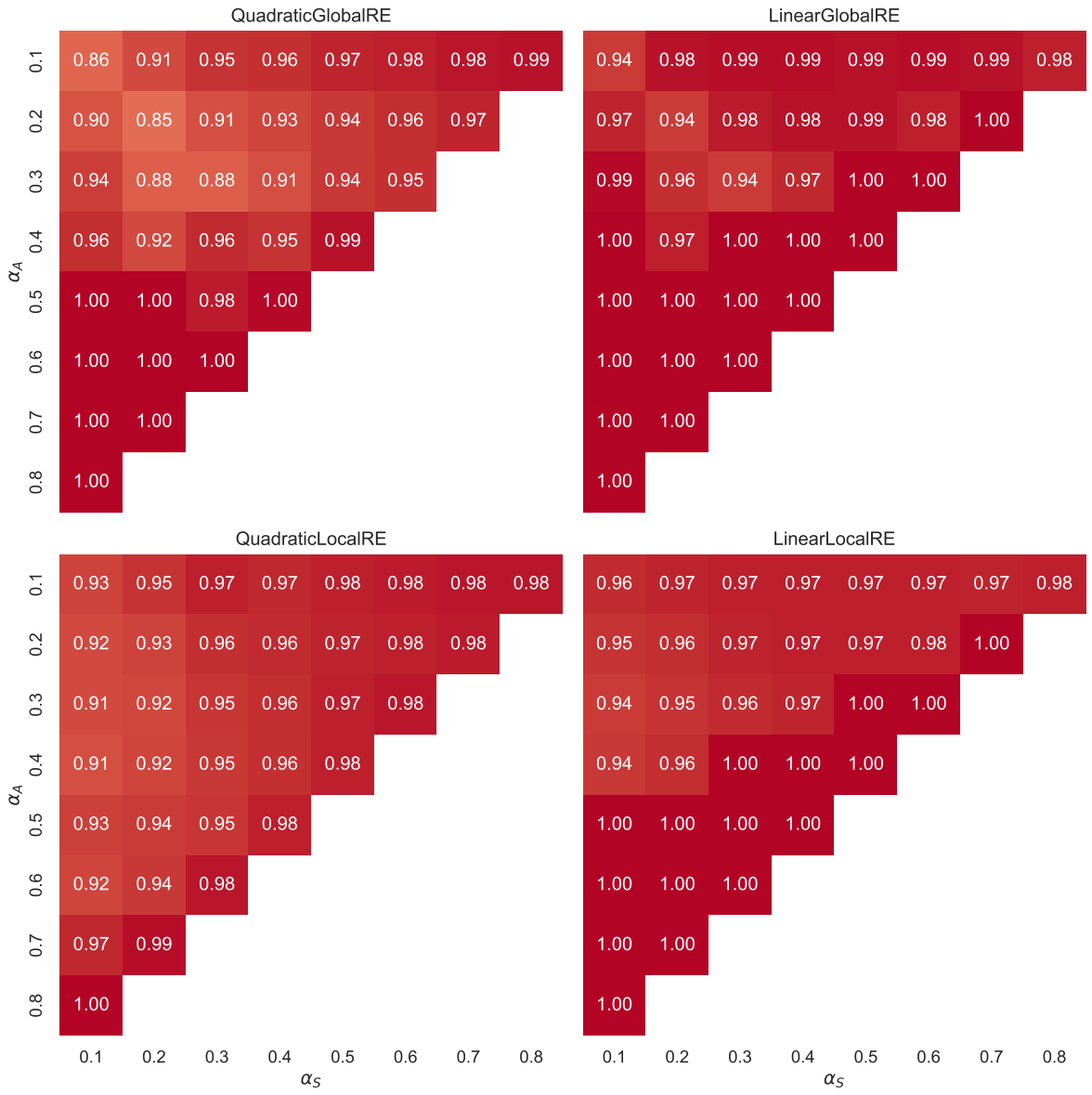


Figure 6.23: Relative share of consistency preserving cases among fixed points (result perspective) grouped by model variant and configuration of weights.

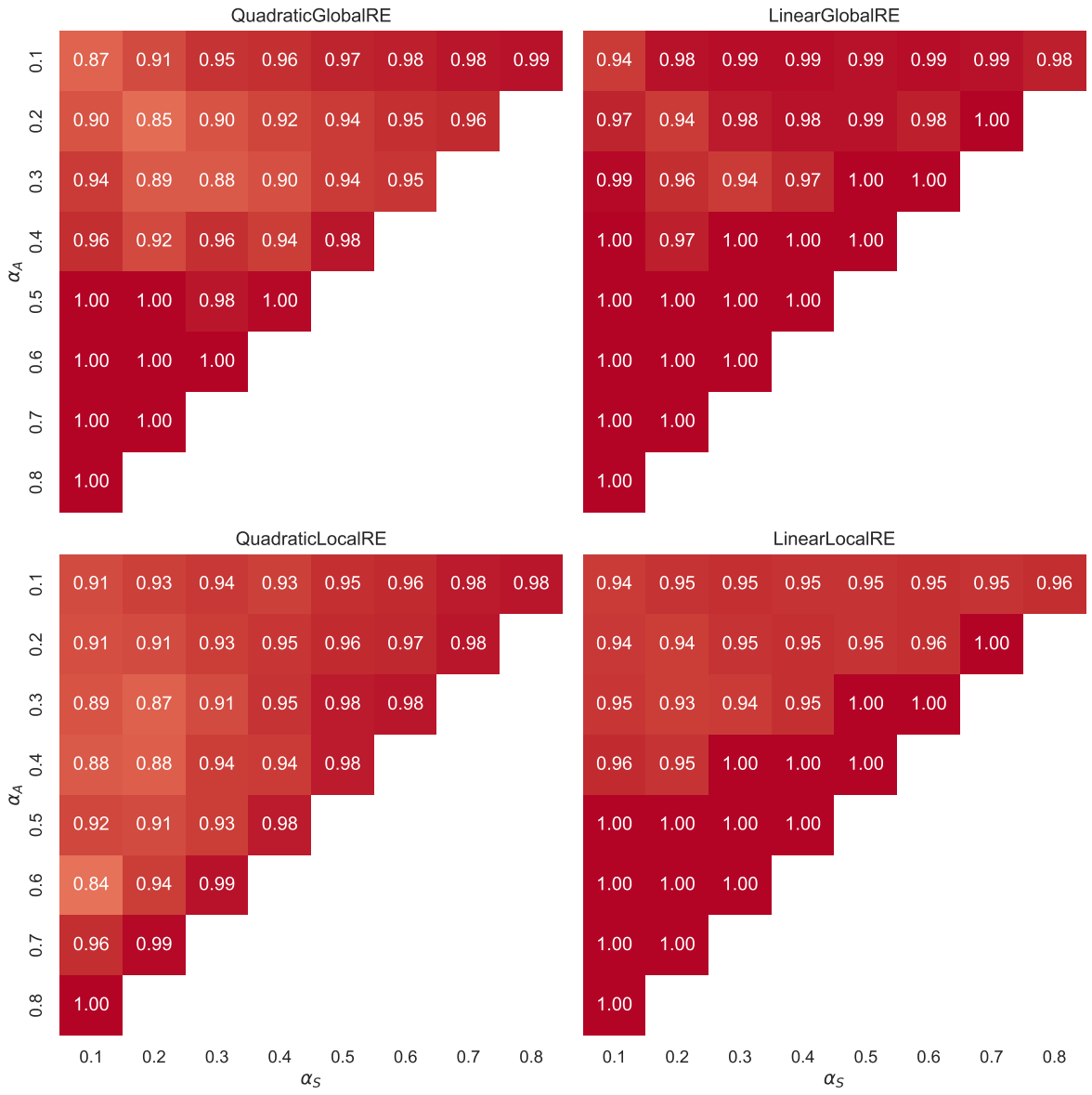


Figure 6.24: Relative share of consistency preserving cases among fixed points (process perspective) grouped by model variant and configuration of weights.

Model	Relative share of global optima with a consistent union	Number of global optima with a consistent union	Number of global optima with consistent commitments
QuadraticGlobalRE	0.931	492856	529359
LinearGlobalRE	0.966	522055	540556
QuadraticLocalRE	0.932	489618	525490
LinearLocalRE	0.966	535532	554525

Table 6.8: Relative share of global optima with a consistent union of commitments and theory

Model	Relative share of fixed points with a consistent union	Number of fixed points with a consistent union	Number of fixed points with consistent commitments
QuadraticGlobalRE	0.915	305081	333436
LinearGlobalRE	0.96	218022	227000
QuadraticLocalRE	0.893	361422	404941
LinearLocalRE	0.973	182164	187163

Table 6.9: Relative share of fixed points (result perspective) with a consistent union of commitments and theory

Model	Relative share of fixed points with a consistent union	Number of fixed points with a consistent union	Number of fixed points with consistent commitments
QuadraticGlobalRE	0.908	340059	374476
LinearGlobalRE	0.96	218065	227097
QuadraticLocalRE	0.911	1333612	1463131
LinearLocalRE	0.994	1233142	1240692

Table 6.10: Relative share of fixed points (process perspective) with a consistent union of commitments and theory

Observations

- The relative shares of consistent unions of commitments and theory among outputs with consistent commitments is very high for all model variants and output types.

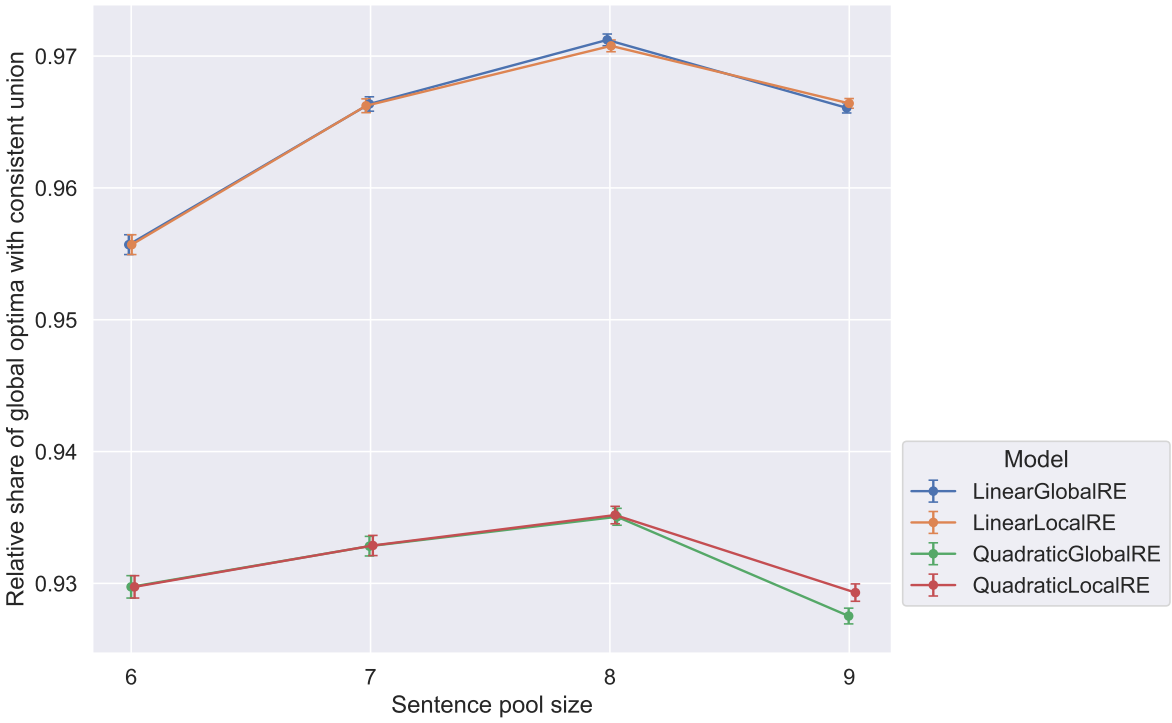


Figure 6.25: Relative share of global optima with a consistent union of commitments and theory grouped by model variant and sentence pool size

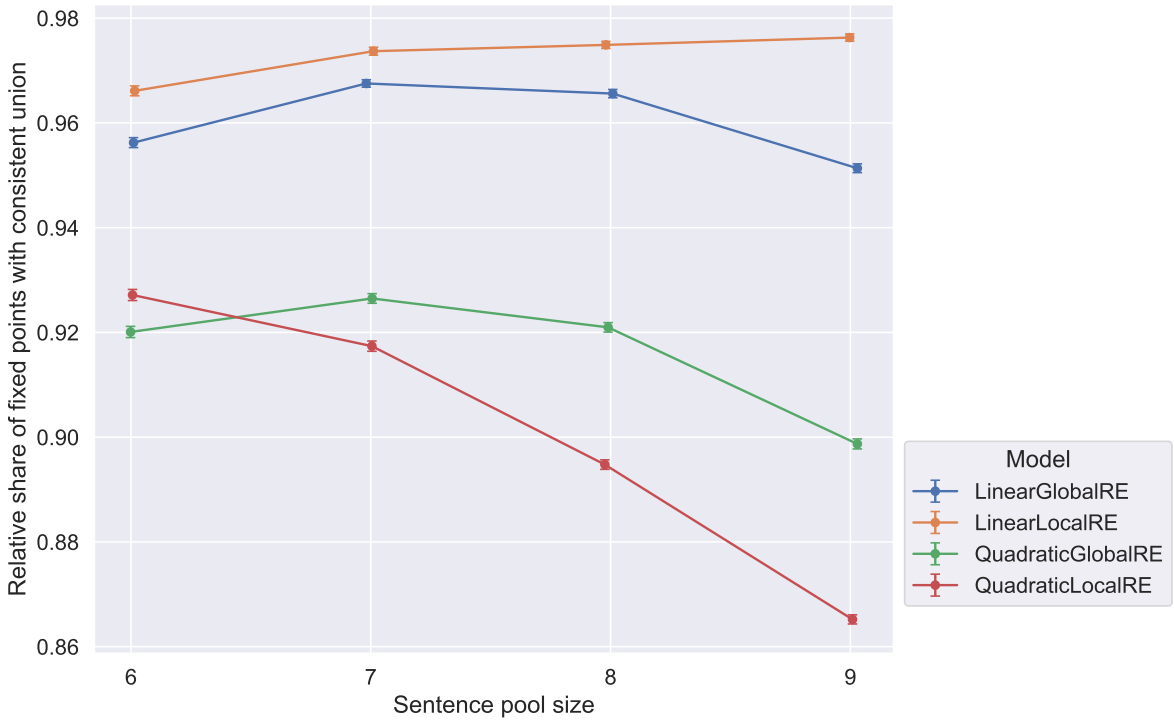


Figure 6.26: Relative share of fixed points (result perspective) with a consistent union of commitments and theory grouped by model variant and sentence pool size

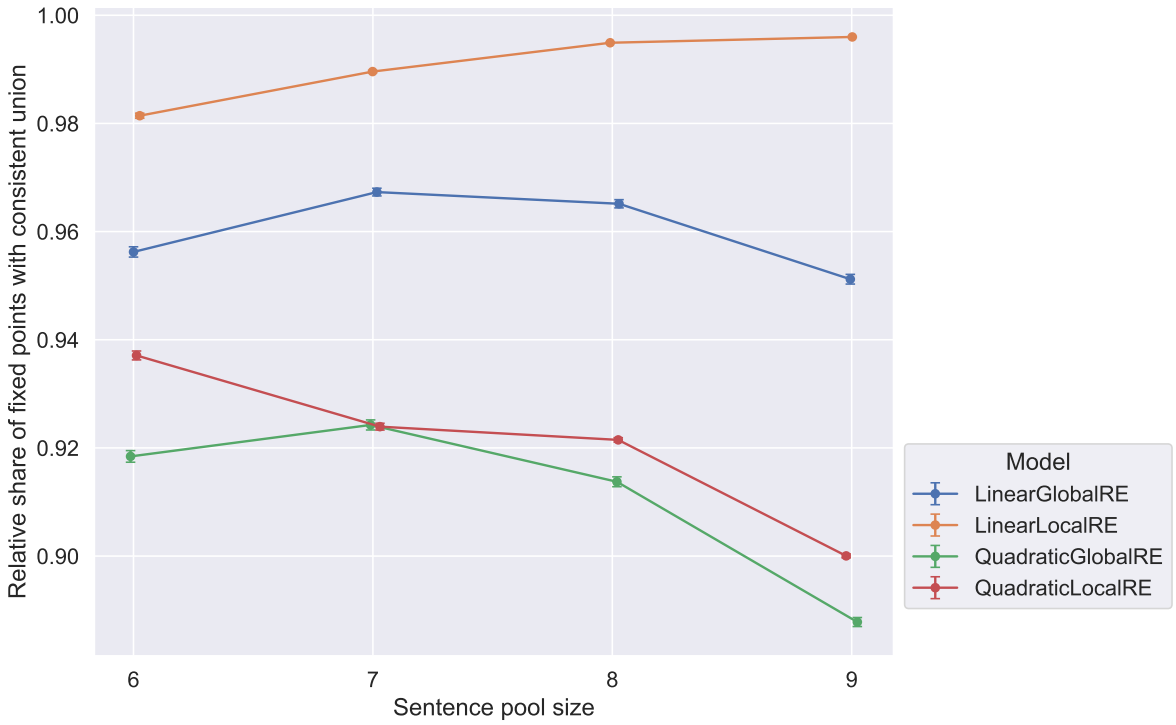


Figure 6.27: Relative share of fixed points (process perspective) with a consistent union of commitments and theory grouped by model variant and sentence pool size

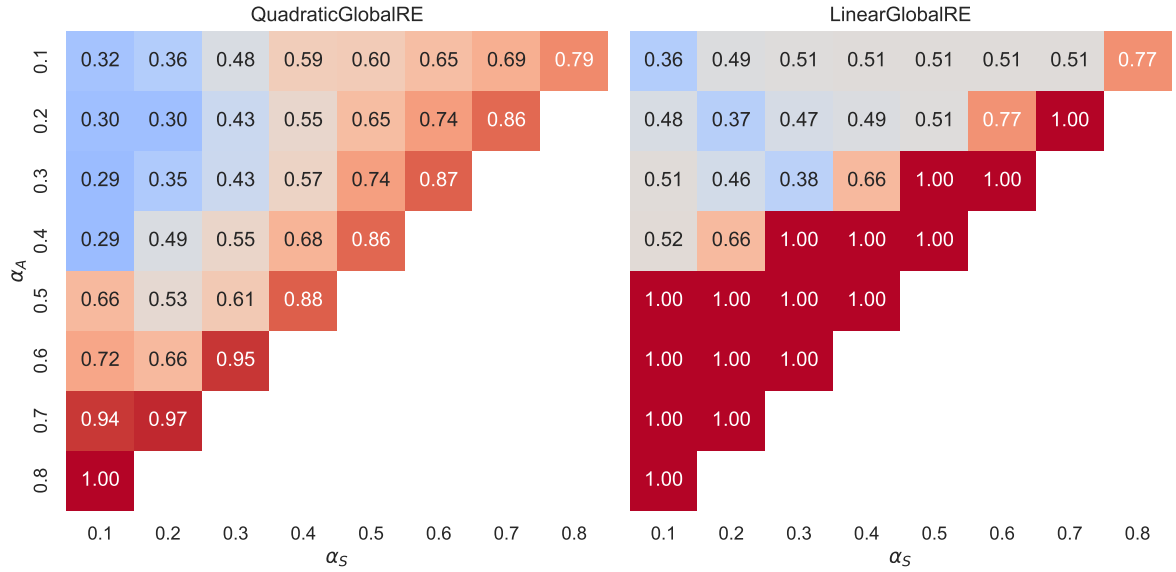


Figure 6.28: Relative share of global optima with a consistent union of commitments and theory grouped by model variant and configuration of weights

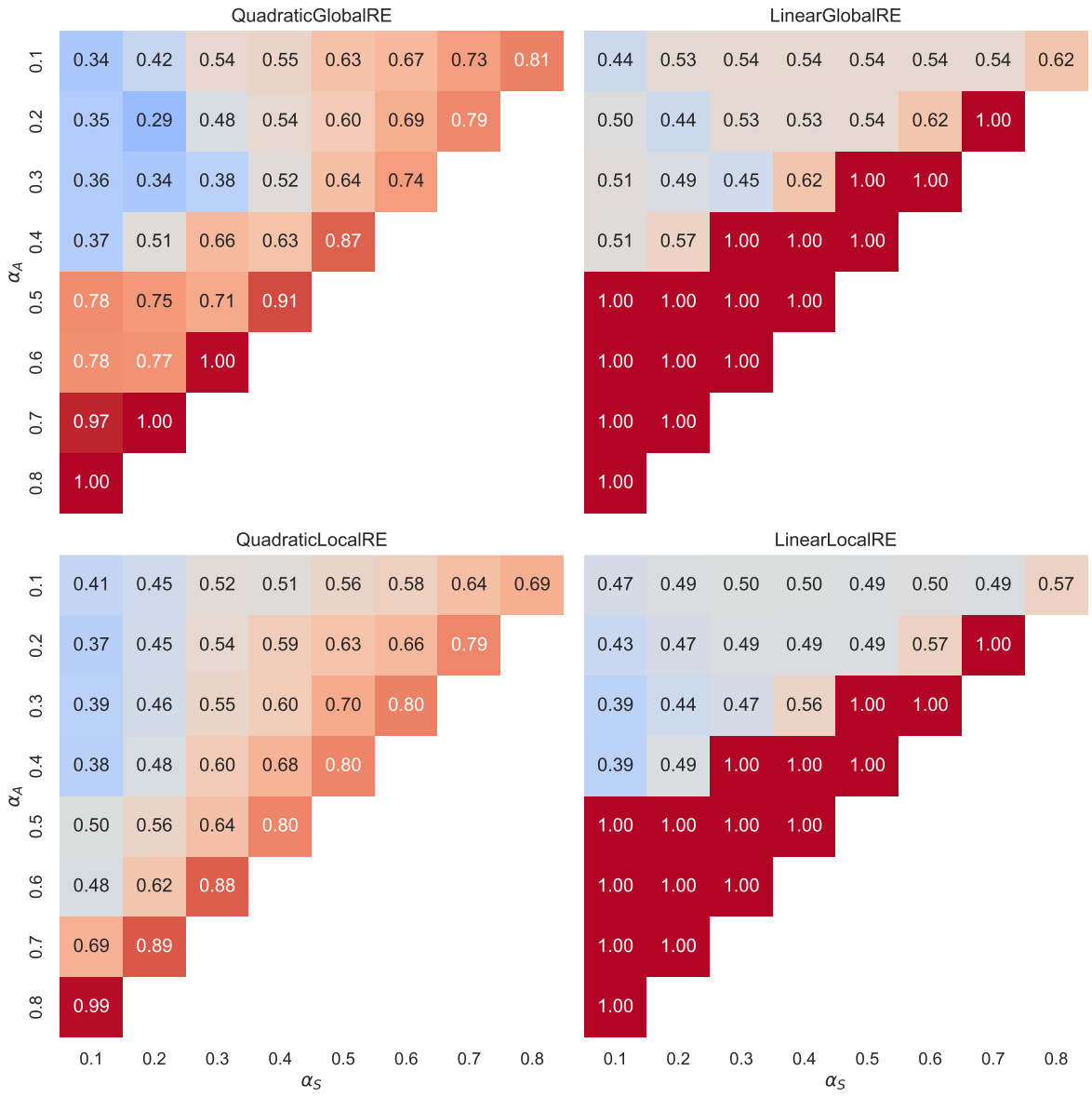


Figure 6.29: Relative share of fixed points (result perspective) with a consistent union of commitments and theory grouped by model variant and configuration of weights

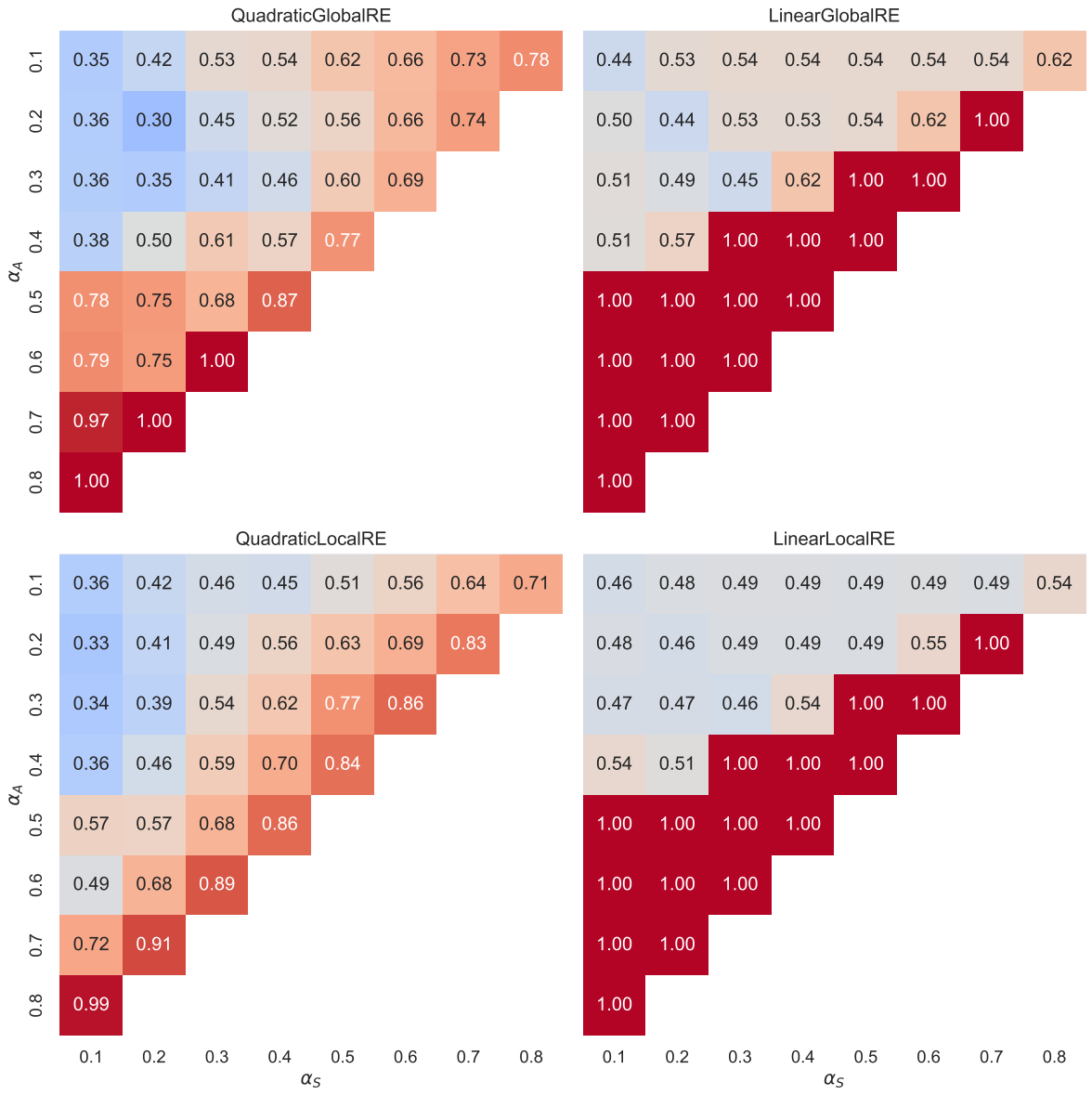


Figure 6.30: Relative share of fixed points (process perspective) with a consistent union of commitments and theory grouped by model variant and configuration of weights

6.2.3.2 Results Grouped by Sentence Pool Size

6.2.3.3 Results Grouped by Configuration of Weights

6.3 Conclusion

Overall, the present ensemble study concerning the three perspectives on the consistency of outputs of RE simulations provides positive results with respect to model variation. The overall relative shares of consistent outputs, inconsistency-eliminating and consistency-preserving cases, as well as consistent unions are satisfactorily high for all model variants.

According to analysing the results further with respect to the sentence pool size, `LinearLocalRE` seems to have the edge over the other model variants in view of increasing sentence pool sizes. Nonetheless, the severely restricted sample that forms the basis of this report would make an extrapolation to even larger sentence pool sizes a highly speculative matter. Further research in this direction is required.

In the more fine-grained analysis according to weigh configurations, we can observe regions of weight configurations that yield desirable behaviour. Moreover, these regions are robust across model variants. This provides at least some motivation to prefer some configurations over others. In particular, it is beneficial to consistency considerations if $\alpha_A > \alpha_F$.

There is a notable difference between quadratic and linear model variants (smooth transitions vs. tipping line), but on its own, this does not serve as a criterion to prefer some model variants over others. See the Appendix A for a presentation of analytical results that explain why linear model variants exhibit tipping lines.

7 Extreme Values for Account, Systematicity, and Faithfulness

7.1 Background

In this chapter, we examine the conditions under which the desiderata account (A), systematicity (S) and faithfulness (F) yield extreme value (i.e., 0 or 1).

Maximal account ($A(\mathcal{C}, \mathcal{T}) = 1$) means that the theory $\mathcal{T}$ fully and exclusively accounts for the commitments $\mathcal{C}$. Full and exclusive account is a condition for full RE states. Conversely, $A(\mathcal{C}, \mathcal{T}) = 0$ holds if a theory completely fails to account for commitments—that is, if for every sentence in the commitments, the theory's closure does not contain this sentence.

The measure of systematicity for a theory $\mathcal{T}$ is defined as follows:

$$S(\mathcal{T}) = G\left(\frac{|\mathcal{T}| - 1}{|\overline{\mathcal{T}}|}\right)$$

with $G = 1 - x^2$ for quadratic models and $G = 1 - x$ for linear models.

Hence, $S(\mathcal{T}) = 1$ if and only if $|\mathcal{T}| = 1$ (i.e., if and only if $\mathcal{T}$ is a singleton theory, e.g., $\mathcal{T} = \{s\}$). Note that it does not matter whether G is linear or quadratic. Furthermore, we have $S(T) = 0$ if and only if $\mathcal{T} = \emptyset$ by definition.

$F(\mathcal{C}|\mathcal{C}_0) = 1$ holds if and only if the initial commitments $\mathcal{C}_0$ are a subset of the commitments $\mathcal{C}$ (expansions of the initial commitments are not penalized). $F(\mathcal{C}|\mathcal{C}_0)$ attains the minimal value of 0 if every sentence of the initial commitments $\mathcal{C}_0$ is missing in or contradicted by the commitments $\mathcal{C}$.

7.2 Results

> **Note**
>
> The results of this chapter can be reproduced with the following Jupyter notebook: https://github.com/debatelab/re-technical-report/blob/main/notebooks/data_analysis_chapter_extreme_values.ipynb.

7.2.1 Overall Results

7.2.1.1 Minimal Values

There is no simulation setup that resulted in a global optimum or a fixed point with a minimal value for account, systematicity or faithfulness. Consequently, we can exclude the consideration of minimal values from the subsequent analysis.

This is a desirable result, as minimal values for A, F and S would constitute quite strange behaviour of the model variants, at least in the range of weights we considered in this study, for we omitted α-weight combinations with zero-valued α weights. Take, for instance, faithfulness: $F(\mathcal{C} \,|\, \mathcal{C}_0) = 0$ would mean that an agent completely departed from their initial commitments $\mathcal{C}_0$, which could be interpreted as changing the subject matter. To the extent that faithfulness matters to some degree (i.e., $\alpha_F \neq 0$), we expect that fixed points and global optima take faithfulness into account (in the sense of $F(\mathcal{C} \,|\, \mathcal{C}_0) \neq 0$ for fixed point commitments or global optima commitments respectively).

7.2.1.2 Maximal Values

Model	Relative share of global optima with maximal account	Number of global optima with maximal account	Number of global optima	Relative share of global optima with maximal systematicity	Number of global optima with maximal systematicity	Relative share of global optima with maximal faithfulness	Number of global optima with maximal faithfulness
QRE	0.115	82318	714584	0.727	519496	0.115	82133
LRE	0.275	192559	700830	0.875	613282	0.288	201631

Table 7.1: Absolute and relative numbers of global optima maximizing various desiderata measures.

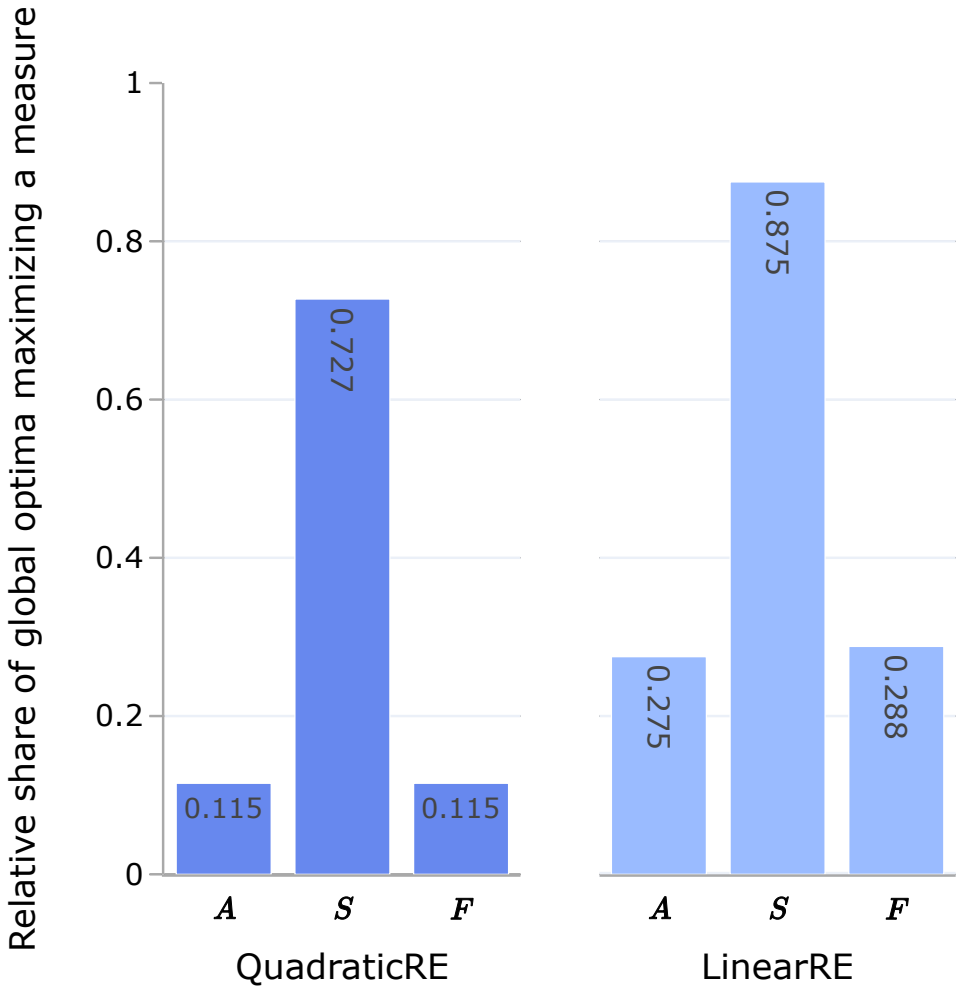


Figure 7.1: Relative shares of global optima maximizing the desiderata measures for account, systematicity and faithfulness

Model	Relative share of fixed points with maximal account	Number of fixed points with maximal account	Number of fixed points	Relative share of fixed points with maximal systematicity	Number of fixed points with maximal systematicity	Relative share of fixed points with maximal faithfulness	Number of fixed points with maximal faithfulness
QGRE	0.166	75903	458147	0.582	266761	0.199	91150
LGRE	0.382	119569	312783	0.724	226486	0.573	179208
QLRE	0.138	81396	588236	0.495	291113	0.083	49095
LLRE	0.639	145846	228122	0.503	114636	0.357	81451

Table 7.2: Absolute and relative numbers of fixed points (resutlt perspective) maximizing various desiderata measures.

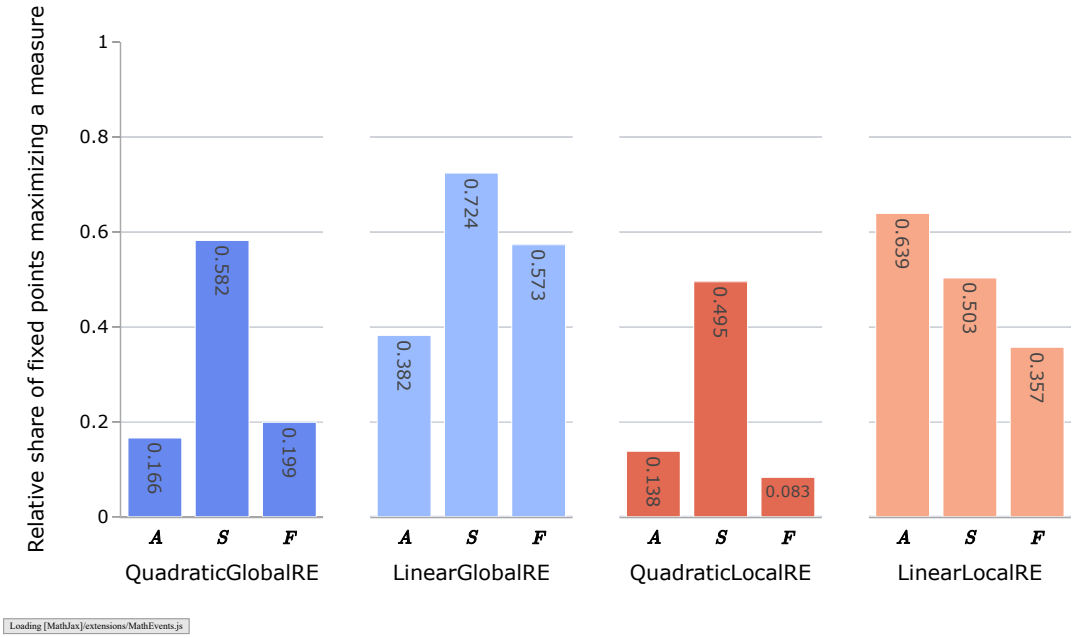


Figure 7.2: Relative shares of unique fixed points (result perspective) maximizing the desiderata measures for account, systematicity and faithfulness

Observations

- Outputs of linear model variants maximize the measures more often than the outcomes

of quadratic models.

- Outputs of all model variants maximize the measure of systematicity more often than the measures for account or faithfulness, excepting fixed points from `LinearLocalRE` (Figure 7.2).
  - It may be easier to maximize S due to the fact that the measure does discriminate singleton theories on the basis of their scope ($|\bar{\mathcal{T}}|$). Thus, there may be many cases in which at least somewhat attractive singleton theories significantly shape the subsequent process of adjustments or the outcome of global optimization.

7.2.2 Results Grouped by Sentence Pool Size

7.2.2.1 Account

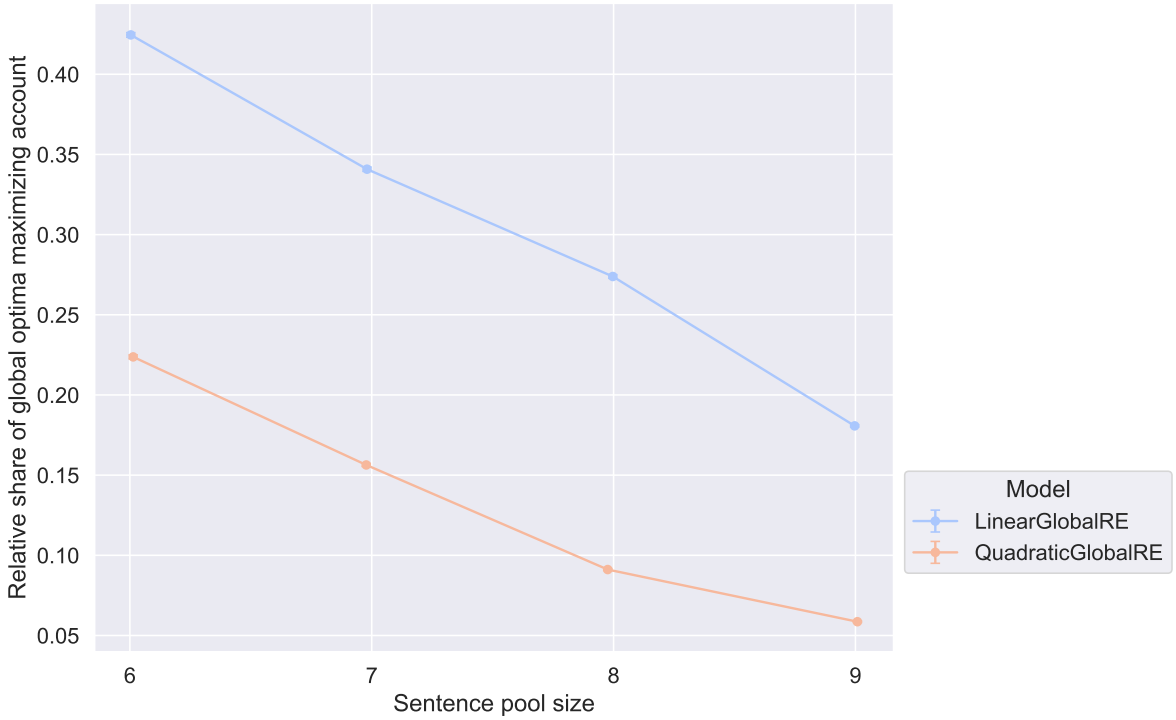


Figure 7.3: Relative share of global optima maximizing the measure for account grouped by model variant and sentence pool size.

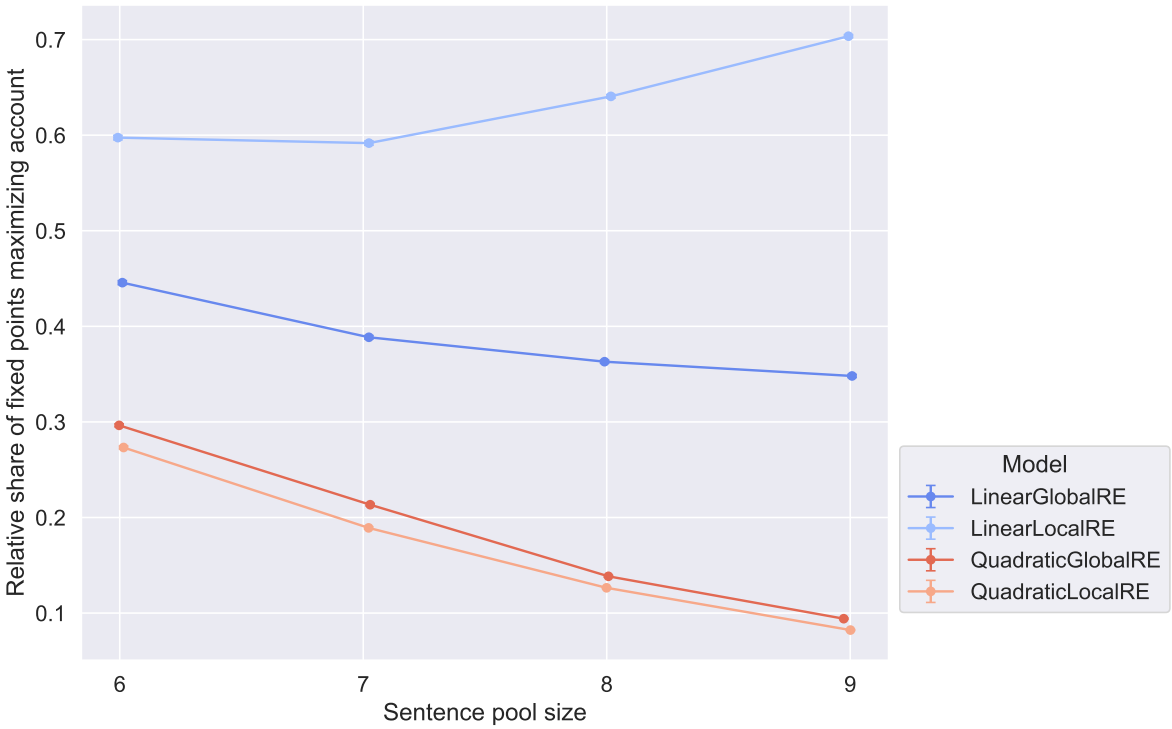


Figure 7.4: Relative share of fixed points (result perspective) maximizing the measure for account grouped by model variant and sentence pool size.

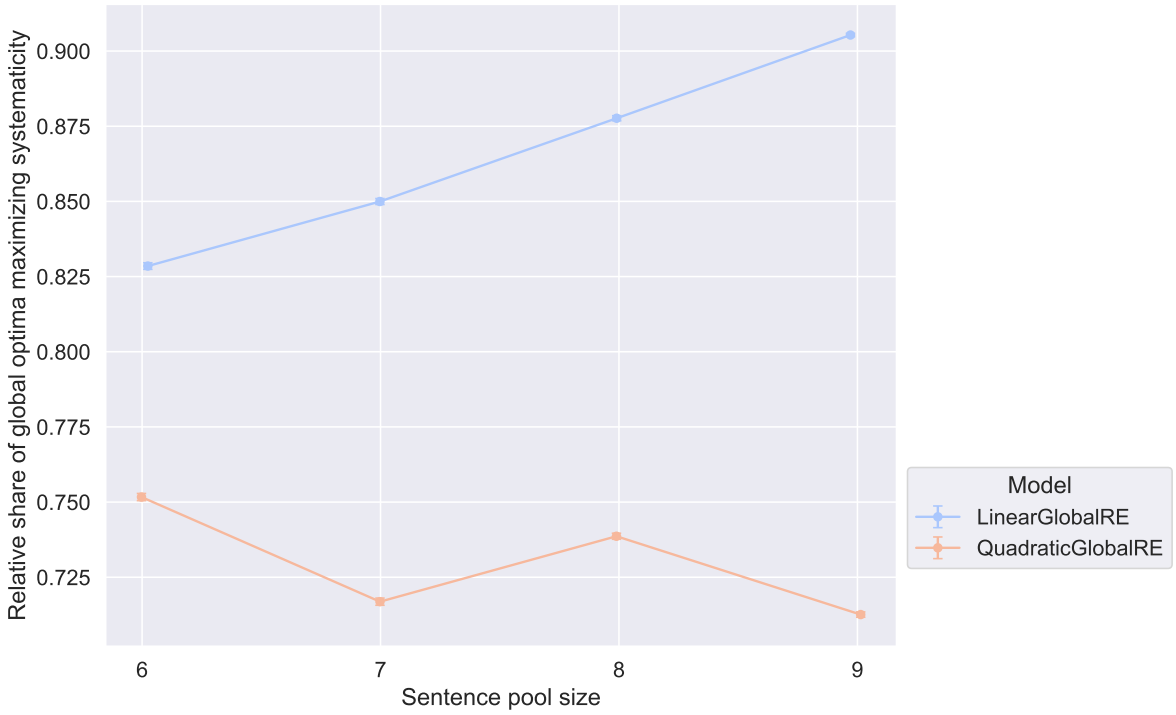


Figure 7.5: Relative share of global optima maximizing the measure for systematicity grouped by model variant and sentence pool size.

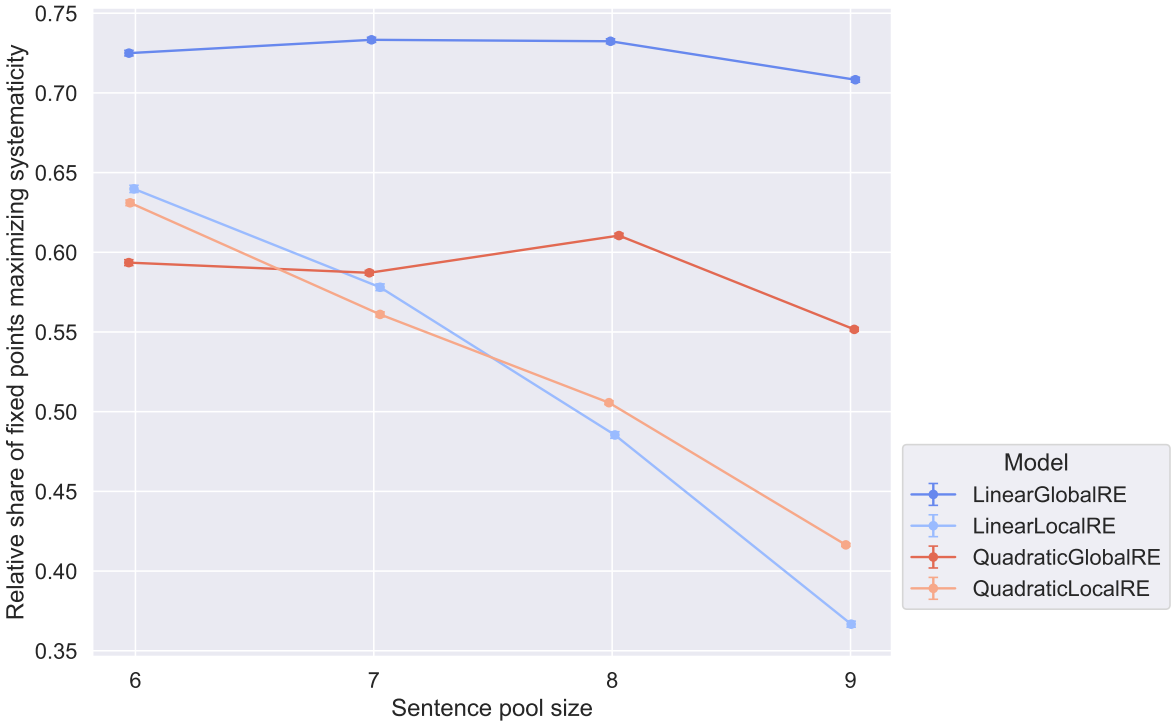


Figure 7.6: Relative share of fixed points (result perspective) maximizing the measure for systematicity grouped by model variant and sentence pool size.

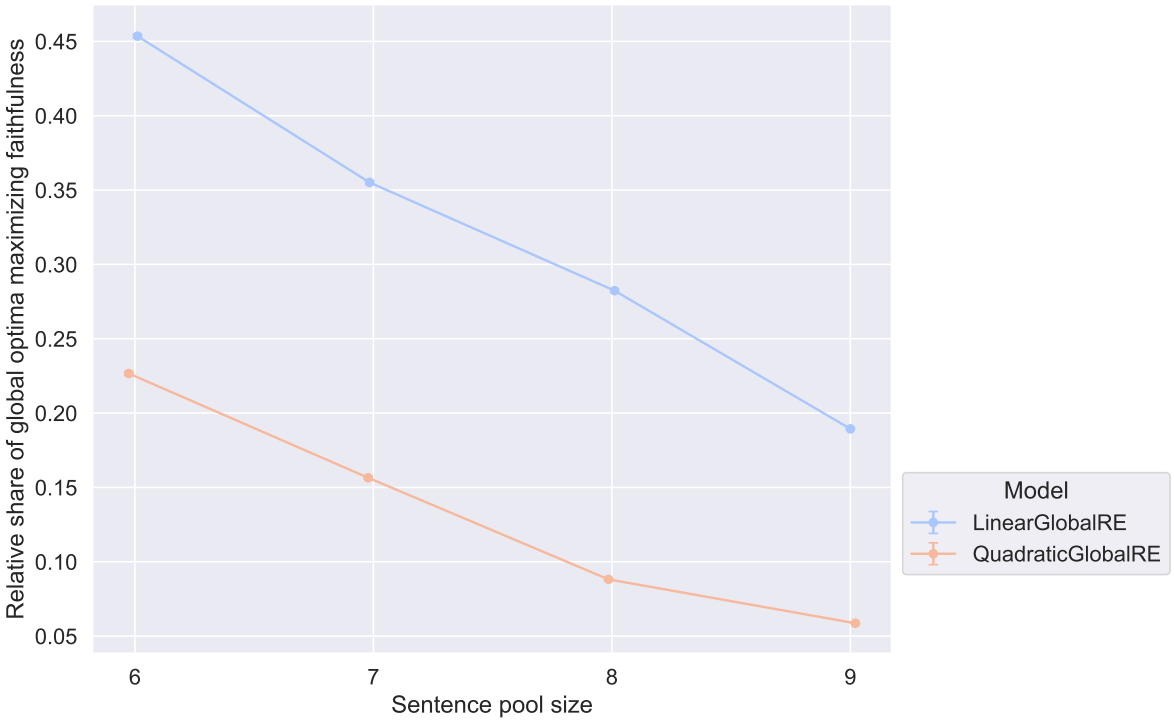


Figure 7.7: Relative share of global optima maximizing the measure for faithfulness grouped by model variant and sentence pool size.

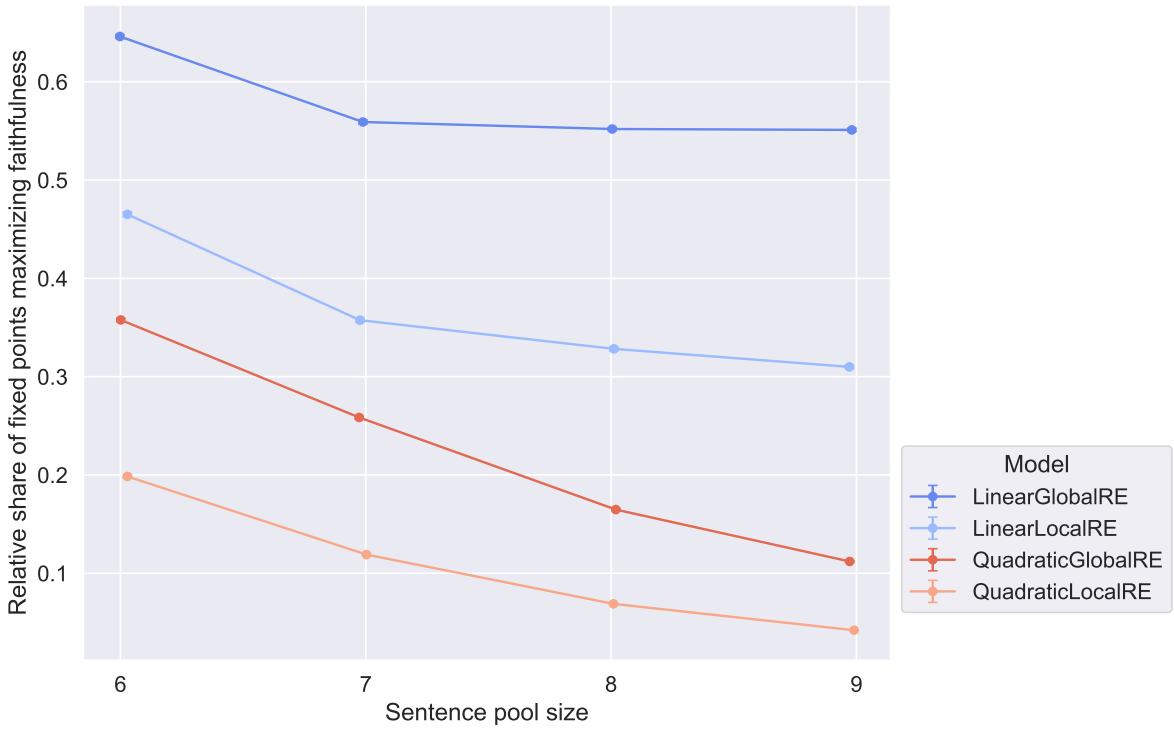


Figure 7.8: Relative share of fixed points (result perspective) maximizing the measure for faithfulness grouped by model variant and sentence pool size.

7.2.2.2 Systematicity

7.2.2.3 Faithfulness

Observations

- The global optima of both quadratic and linear model variants maximize account (Figure 7.3) and faithfulness (Figure 7.7) less frequently for larger sentence pool sizes.
- This tendency is less pronounced for fixed points (result perspective) in Figure 7.4 and Figure 7.8 , respectively.
- The relative share of fixed points (result perspective) that maximize systematcity is not affected by the sentence pool size for global model variants (Figure 7.6). In contrast this relative share decreases with increasing sentence pool sizes for local model variants.

7.2.3 Results Grouped by Configuration of Weights

7.2.3.1 Account

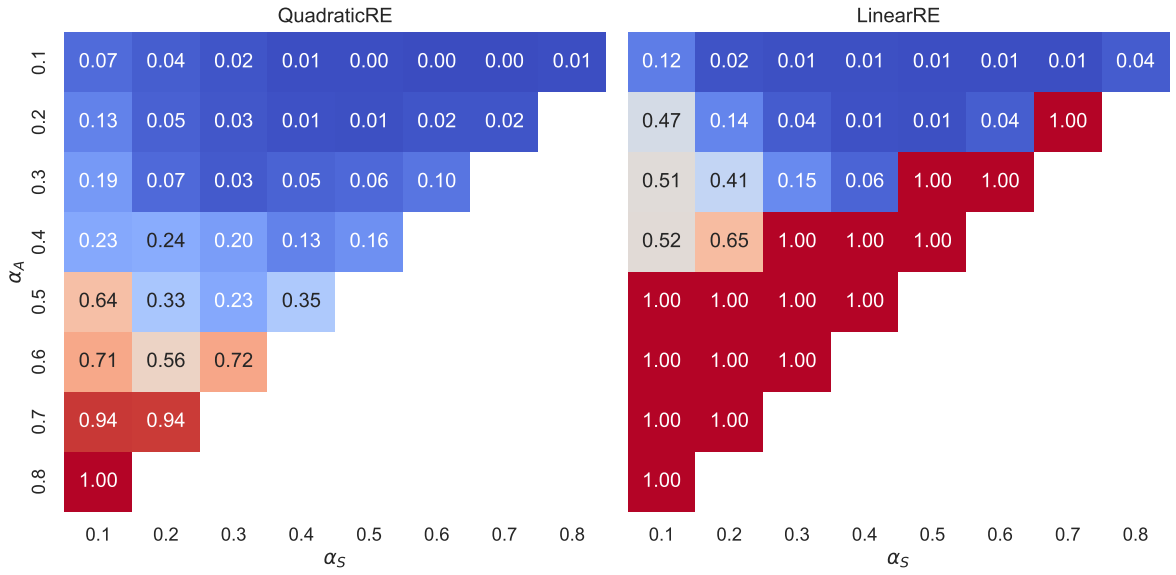


Figure 7.9: Relative share of global optima maximizing the measure for account grouped by model variant and configuration of weights.

Observation

- Linear model variants exhibit a "tipping line". For $\alpha_A > \alpha_F$ global optima and fixed points always maximize the measure for account. For an explanation, see Appendix A.

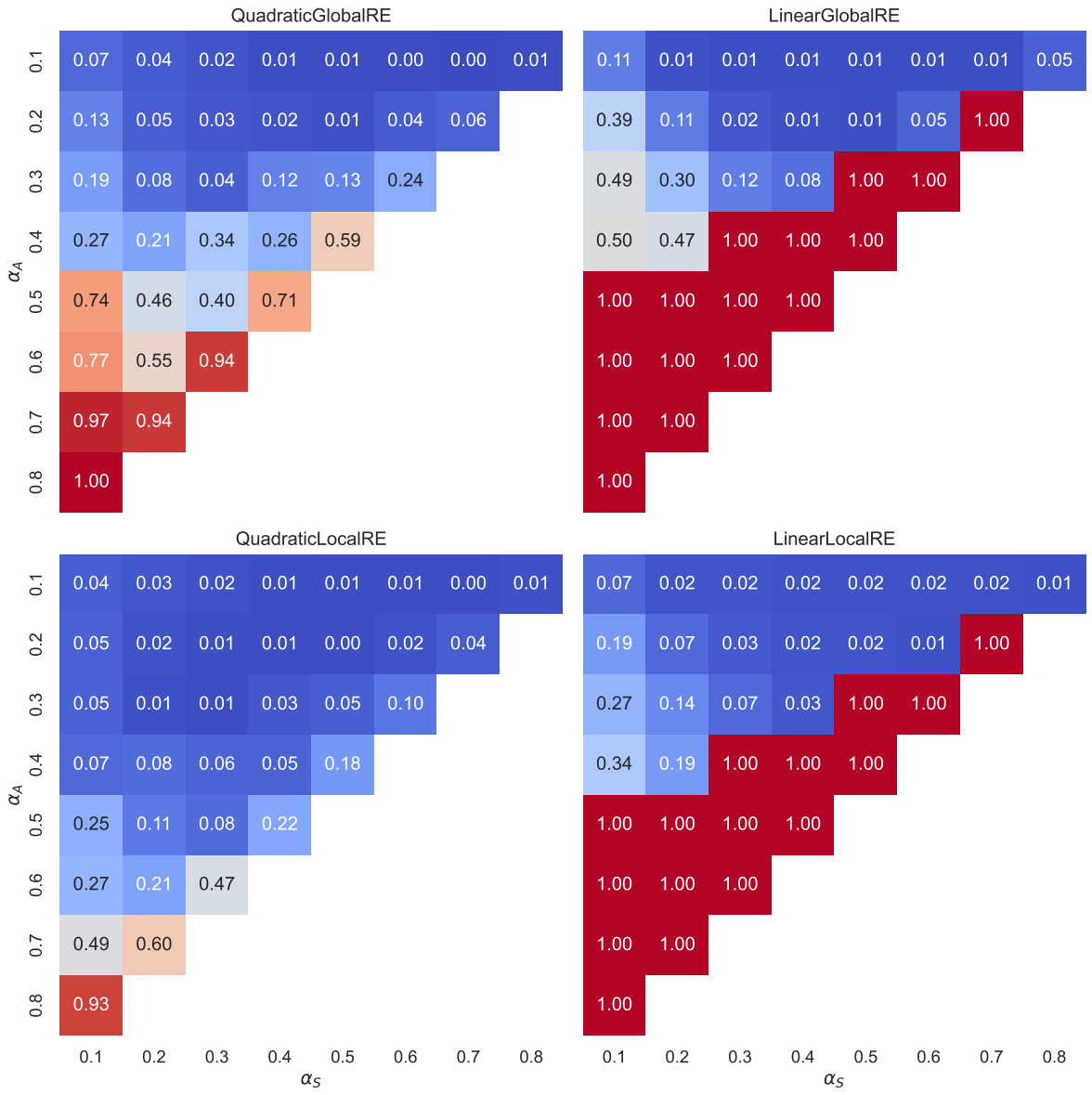


Figure 7.10: Relative share of fixed points (result perspective) maximizing the measure for account grouped by model variant and configuration of weights.

- Quadratic model variants exhibit a gradient with increasing relative shares for higher values of α_A.

7.2.3.2 Systematicity

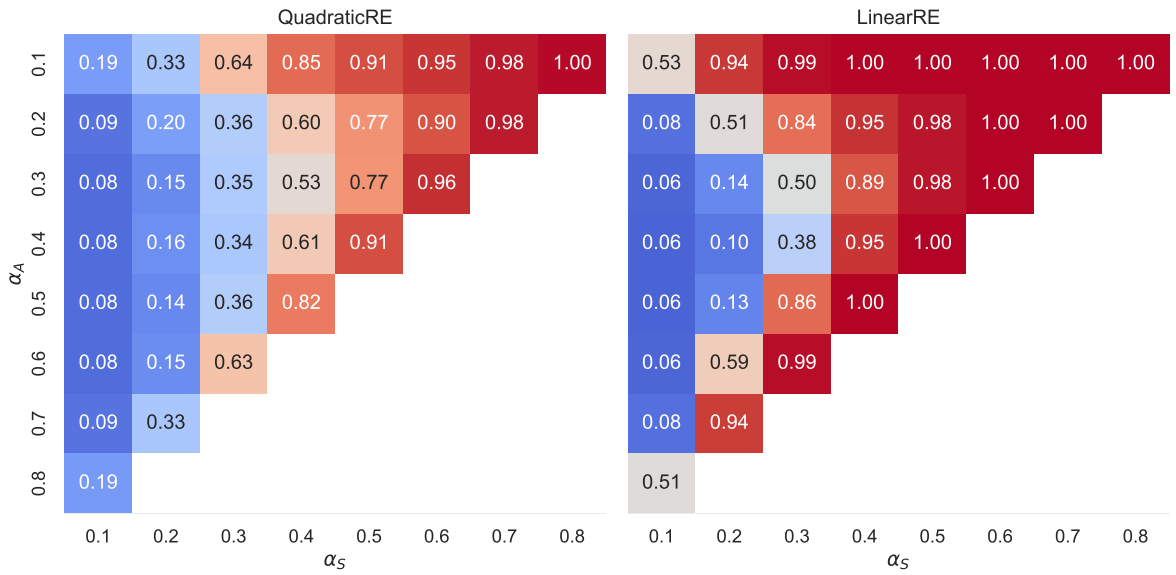


Figure 7.11: Relative share of global optima maximizing the measure for systematicity grouped by model variant and configuration of weights.

Observations

- For all model variants and outputs, we can observe a gradient of increasing relative shares of outputs with maximal systematicity for increasing values of α_S.
- Moreover, the relative share also increases for decreasing weights for α_A. If account does not receive much weight, the theory can be optimized with respect to systematicity more independently of the commitments, even if α_S is low.

7.2.3.3 Faithfulness

Observations

- Linear model variants exhibit a "tipping line". For $\alpha_F > \alpha_A$ global optima and fixed points always maximize the measure for faithfulness. For an explanation, see Appendix A.
- Quadratic model variants exhibit a gradient with increasing relative shares for higher values of α_F.

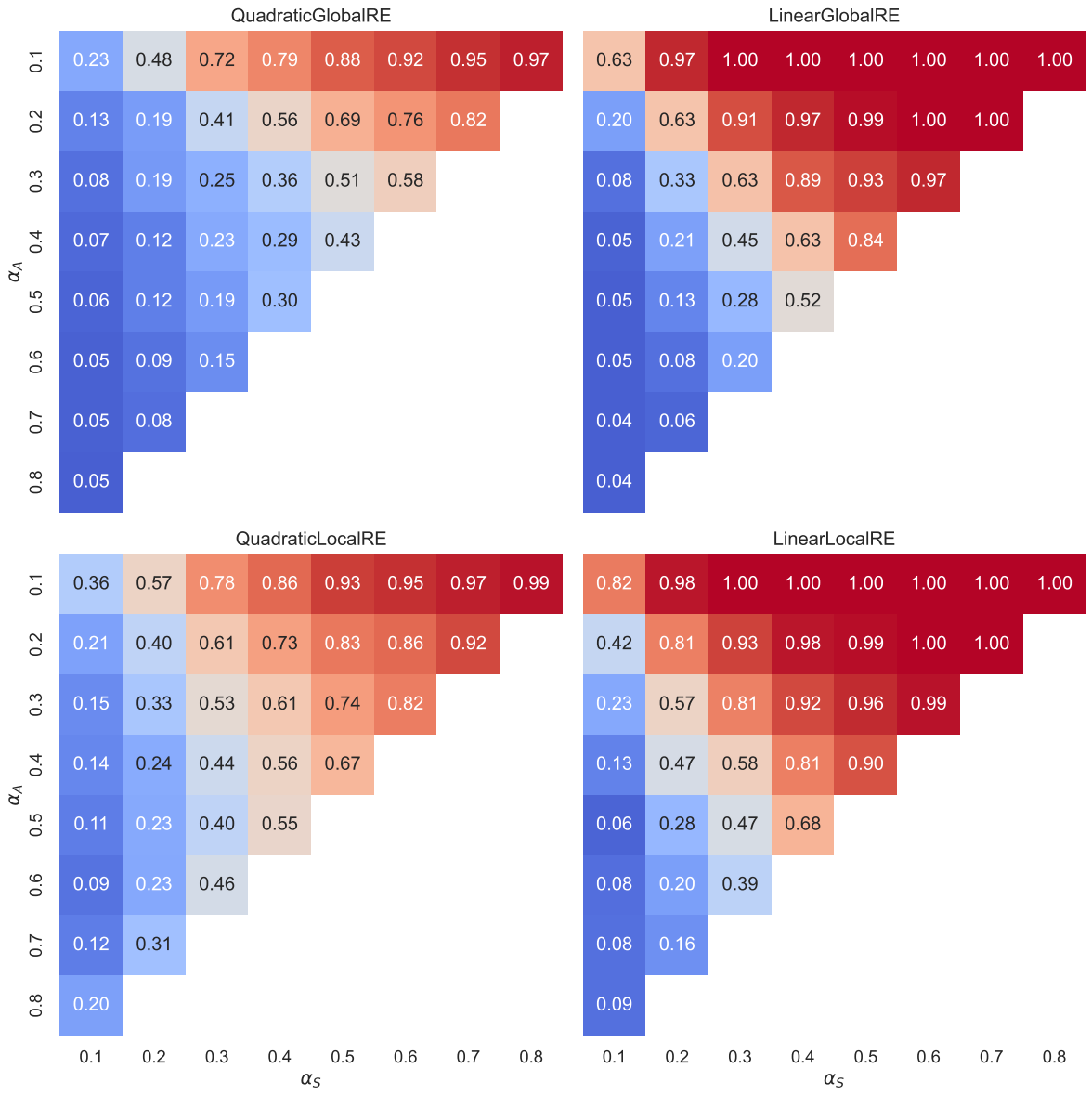


Figure 7.12: Relative share of fixed points (result perspective) maximizing the measure for systematicity grouped by model variant and configuration of weights.

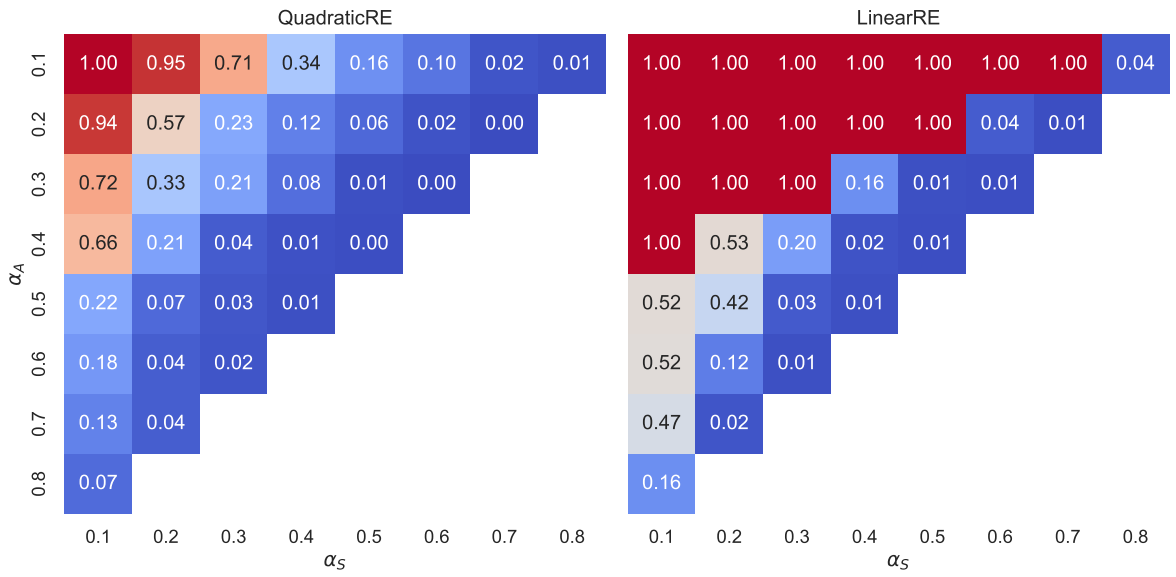


Figure 7.13: Relative share of global optima maximizing the measure for faithfulness grouped by model variant and configuration of weights.

7.3 Conclusion

Many observations in this chapter are not surprising. It is to be expected that increasing the weight results in higher relatives shares of maximized measures. Nonetheless, this is a reassuring result from the viewpoint of model evaluation, indcating that configuring weights has forseeable consequences.

The high relative shares of outputs maximizing the measure for systematicity may be a consequence of a shortcoming in the measure for systematicity. If $|\mathcal{T}| = 1$, then $S(\mathcal{T}) = 1$ irrespective of $|\overline{\mathcal{T}}|$. That is the measure for systematictiy does not discriminate between singleton theories on the basis of their scope ($\overline{\mathcal{T}}$). This renders all singleton theories equally and maximally attractive according to the measure of systematicity. For another consequence of frequently maximizing the measure for systematicity, see Appendix B.

Further exploration is required to provide full explanations for the more salient observations. For example, one could analyze the "evolution" of theories during RE processes.[1] Are singleton theories chosen in the first adjustment step and not altered afterwards? Or do RE processes set out with larger theories and are elements remove subsequently?

[1]This information is already available in the data.

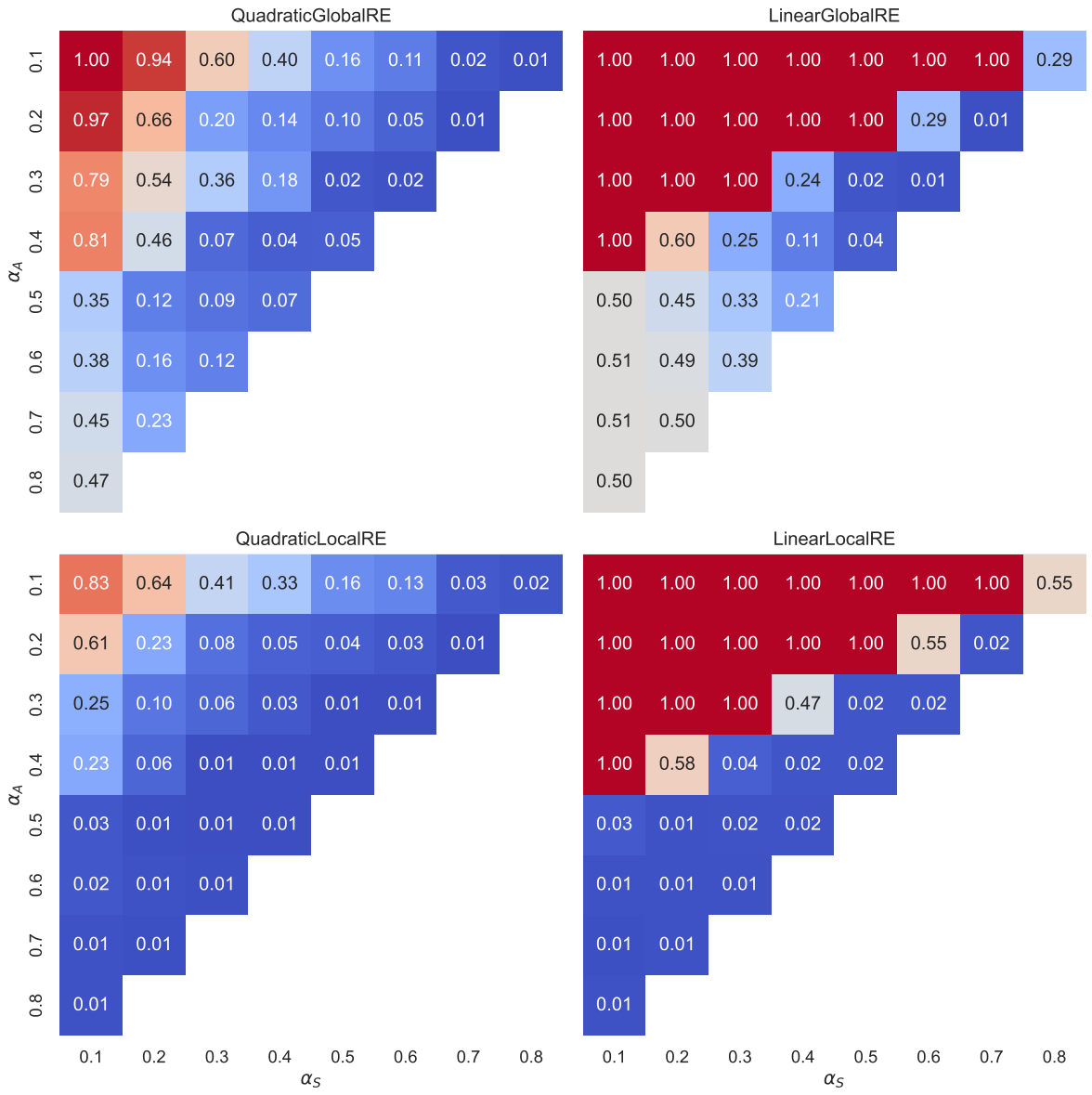


Figure 7.14: Relative share of fixed points (result perspective) maximizing the measure for faithfulness grouped by model variant and configuration of weights.

8 Summary

8.1 Overview

This report thoroughly assessed the formal RE model by Beisbart, Betz, and Brun (2021) by numerical investigation. We ran computer simulations for a broad spectrum of model parameters and initial conditions and used four different model variants. In this chapter, we summarize the most important findings with respect to the metrics described in Section 2.3.

Global Optima and Fixed Points

In Chapter 4, we investigated whether fixed points are global optima (GO efficiency) and, conversely, whether global optima are reachable by equilibration processes (GO reachability).

- Overall, GO efficiency is high for semi-globally optimizing models and medium-high for locally optimizing models.
- GO efficiency drops for locally optimizing models with the size of the sentence pool.
- For $\alpha_A < \alpha_S$, GO efficiency of the `LinearLocalRE` model is as high as of the models `QuadraticGlobalRE` and `LinearGlobalRE`.
- GO reachability is low to medium for all models.
- All models except the `QuadraticGlobalRE` model perform worse concerning GO reachability with an increase in the size of the sentence pool.
- The `QuadraticGlobalRE` model outperforms all other models on average.
- The`LinearLocalRE` model reaches a higher GO efficiency than the `QuadraticLocalRE` model, but it is the other way around with respect to GO reachability.

Full RE States

In Chapter 5, we explored whether fixed points and global optima attain full RE states (i.e., global optima for which the theory fully and exclusively accounts for the commitments).

- Overall, the relative share of full RE states among global optima and fixed points is rather low.
- Heatmaps reveal combinations of weights for `GlobalQuadraticRE`, `GlobalLinearRE` and `LinearLocalRE`, where the relative share of full RE states among the outputs is acceptable.
- There is a slight negative trend for the relative shares of full RE states among global optima and fixed points (result perspective) for increasing sentence pool sizes.
- The sentence pool size does not affect the relative share of full RE fixed points (process perspective) of`LinearLocalRE`.

Consistency

In Chapter 6, we assessed different aspects of consistency conduciveness of the model variants.

- The overall relative shares of consistent outputs, inconsistency-eliminating and consistency-preserving cases, as well as consistent unions, are satisfactorily high for all model variants.
- In view of increasing sentence pool sizes, `LinearLocalRE` performs best with respect to all examined aspects of consistency.
- There are regions of weight configurations ($\alpha_A > \alpha_F$) that yield desirable behaviour concerning consistency across all model variants.
- A salient "tipping line" in heatmaps of linear model variants marks off regions of weight configurations that yield a fundamentally different behaviour. The analytical results from Appendix A explain these observations.

Extreme Measure Values

In Chapter 7, we investigated whether global optima and fixed points yield extreme values in the normalized measures A, F and S.

- Overall, there are no surprising observations: Increasing the weight of a specific measure leads to more outputs that maximize the corresponding measure.

8.2 Appendices

The appendices include additional material, which can be used to explain some of the simulation results and which motivates suggestions for further research.

The Tipping Line of Linear Model Variants

In Appendix A, we provide analytical results concerning a "tipping line" in linear model variants that help to explain various observations in the report.

- For $\alpha_A > \alpha_F$, global optima of linear model variants always achieve full and exclusive account ($A(\mathcal{C}, \mathcal{T}) = 1$).
- For $\alpha_F > \alpha_A$, the commitments of global optima of linear model variants are always maximally faithful to the initial commitments ($F(\mathcal{C} \,|\, \mathcal{C}_0) = 1$).
- These results can be generalized to fixed points of the linear model variants.

Note that the "tipping-line behaviour" we observed in the simulation results for the linear model variants concern their performance with respect to the various validation metrics and not which global optima and fixed points are reached. In other words, in each of the two regions ($\alpha_A > \alpha_F$ and $\alpha_F > \alpha_A$), global optima and fixed points will generally depend on the α-weight combinations. Otherwise, we would have observed the tipping-line behaviour in all results for the linear model variants, which we didn't.

The described restriction of the tipping-line behaviour is essential because, without this restriction, we could formulate a substantive objection against using the linear model variants. If global optima (and fixed points, respectively) would only depend on whether $\alpha_A > \alpha_F$ or $\alpha_F > \alpha_A$, and, accordingly not change within these regions, the model would fail to represent different decisions as how to balance account and faithfulness in reaching reflective equilibria—at least, the decision would be trivialized into a binary decision. However, the whole idea of using the proposed achievement function with α weights on a continuous scale is to allow for a fine-grained spectrum of balancing the different desiderata.

Trivial Endpoints

In Appendix B, we analyzed whether the model variants yield "trivial" outputs—that is, global optima or fixed points that consist of singleton theories and commitments.

- Overall, the relative share of trivial global optima and fixed points (result perspective) is very low for the quadratic model variants.
- Linear model variants exhibit substantially more trivial global optima, but the relative shares are still low.
- `LinearLocalRE` exhibits a substantial share of trivial fixed points from the process perspective but not from the result perspective.
- The relative shares of trivial global optima or fixed points tend to decrease with increasing sentence pool sizes.
- In quadratic model variants, the α weights have only a small impact on the relative shares of trivial endpoints.

Alternative Systematicity Measures

In Appendix C, we motivated alternative systematicity measures in view of shortcomings of the original systematicity measure in Beisbart, Betz, and Brun (2021). We discussed their advantages and disadvantages in terms of various desiderate for such measures (see Table C.1 for an overview).

One sophisiticated systematicity measure (Section C.3.1) is able to satisfy five out of six desiderata, but no proposed measure is able to satisfy all six of them. In view of the only intuitively motivated desiderata and the lack of simulation data, we conclude that these results are preliminary. In particular, they do not prescribe to replace the original measure of systematicity.

8.3 Conclusion

The results we arrived at are insufficient to draw general conclusions about the overall performance of the four analyzed model variants. Neither did we find conclusive evidence to exclude one model as generally inadequate, nor did we identify one model that outperforms the others in all aspects. Instead, each model variant meets some of the validation criteria to a sufficient

degree within some ranges of simulation setups. In cases where a model variant performs poorly on average (over the spectrum of simulation setups), the others did as well. In other words, the performance of a model depends crucially on the specifics of the simulation setup (e.g., the chosen dialectical structure, sentence pool size, α weights and initial commitments) and the evaluation criterion at hand.

This does not mean there are no differences between the model variants. Instead, in a specific context of using the RE model, there might be good reasons to prefer some model variant over the other. This is because the context might fix certain specifics of the simulation setup and provide independent reasons for them. Similarly, the context might give us a more nuanced picture of the relative importance of the different validation criteria. In light of such specifications, the results we presented can be used (possibly in combination with additional analyses) to choose a specific model (or at least exclude some).

For instance, the context might prescribe a limited range of α-weight combinations. In other words, there might be independent reasons of how to balance account, faithfulness and systematicity. We already saw that a model's performance is often highly sensitive to the chosen α weights. Within this region, one might repeat all those dependency analyses we only averaged over all α-weight configurations (e.g., a model's performance in dependence of the sentence pool size). Then, it can (and will) still happen that the models perform differently with respect to the different validation criteria (consistency, reaching global optima and full RE states). However, that only means that there is a trade-off between these metrics. In other words, in addition to balancing account, faithfulness and systematicity, there is a balancing of those desiderata that are connected to the used validation criteria.

From this perspective, it is perhaps not that surprising and worrisome that the described results are mixed but in perfect agreement with central ideas about RE.

8.4 Outlook

In many ways, this technical report is but a starting point for future lines of research. In the following, we describe some promising and pressing issues that call for further research.

Note that the current Python implementation of the model is designed to facilitate extending the model (as demonstrated by the three model variants used in this report). Various components of the formal model, for instance, the measures account, faithfulness, and systematicity can be changed with a few lines of code (source).

8.4.1 The Neighborhood Depth and the Search Strategy of Locally Optimizing Model Variants

The local model variants examine available candidate positions for adjustments during RE processes in a small neighborhood of the current position. For this report, the search depth

was confined to adjusting one single sentence per adjustment step. A particular shortcoming of such small neighborhood depths is that they may "miss" sensible adjustments that involve arguments with more than one premise.[1] In particular, the adjustment of theories might be severely restricted.

It is important to note that considering larger neighborhood depths reintroduces an exponential growth of the search space depending on the size of the sentence pool. One might, therefore, worry that enlarging the neighborhood depth defies the original motivation to use locally optimizing models—namely, providing a model that works computationally feasible with larger sentence pools.

In view of this and additional reasons, it is worthwhile to devise and analyze locally optimizing models that implement other search strategies for finding subsequent epistemic states. For instance, the process might mimic a random walk, or we might allow the model to "backtrack" different branches, enabling them to avoid dead-ends (i.e., mere local optima).

8.4.2 Alternative Systematicity Measures

The measure of systematicity in the original formal model of Beisbart, Betz, and Brun (2021) has a shortcoming, as it does not discriminate between singleton theories on the basis of their scope (for formal details, see Section 7.1).

In Appendix C, we discussed several alternative suggestions to define systematicity and began to analyze them with respect to some intuitive criteria. These preliminary considerations should be complemented with the exploration of simulation results of corresponding model variants.

8.4.3 The Inferential Density of Dialectical Structures

We did not analyze the performance of the model variants in dependence on the inferential density of the randomly generated dialectical structures (for the definition, see Section 2.4.3). One reason for this omission was the worry that the generated 50 dialectical structures per sentence pool hardly correspond to a representative sample of dialectical structures. Accordingly, we did not analyze whether and to what extent model outcomes depend on properties of the dialectical structure other than the sentence pool size. Hence, it may be interesting to treat, for instance, inferential density as an independent variable to gain new insights about the model's behaviour.

[1] Results that might suggest such a shortcoming of local model variants can be found in Figure 4.4 and Figure 4.5.

8.4.4 Extrapolation to Larger Sentence Pools

We considered only a confined range of sentence pools with few sentences (12, 14, 16 and 18). As it stands, the results of this report provide no solid basis to extrapolate our findings to larger sentence pools. Such results are, however, needed since it is pretty clear that applications of the formal RE model to somewhat realistic cases will involve much more sentences.[2] It is, in particular, important to know whether and under which conditions locally optimizing model variants can reach global optima since a semi-global optimization is computationally infeasible with larger sentence pools. In these cases, some form of local optimization has to take over. However, the prospects of using locally optimizing models have to be evaluated carefully beforehand. To arrive at better estimates, one would need dedicated ensembles of simulations comprising larger sentence pools that simultaneously allow the calculation of global optima as reference points.

[2]For instance, the reconstruction of Thomson's famous "The Trolley Problem" (2008) by Rechnitzer (2022) involves 25 (unnegated) sentences. This would amount to the daring task of considering 3^{25} (roughly 850 billion) candidates per commitment adjustment step in an RE process with a semi-globally optimizing model variant.

References

Beisbart, Claus, Gregor Betz, and Georg Brun. 2021. "Making Reflective Equlibrium Precise: A Formal Model." *Ergo* 8 (0). https://doi.org/10.3998/ergo.1152.

Betz, Gregor. 2010. *Theorie dialektischer Strukturen*. Frankfurt am Main: Klostermann.

———. 2013. *Debate Dynamics: How Controversy Improves Our Beliefs*. Synthese Library. Dordrecht: Springer Netherlands.

Freivogel, Andreas. 2023. "Does Reflective Equilibrium Help Us Converge?" *Synthese* 202 (6): 1–22. https://doi.org/10.1007/s11229-023-04375-0.

Rechnitzer, Tanja. 2022. "Turning the Trolley with Reflective Equilibrium." *Synthese* 200 (4): 1–28. https://doi.org/10.1007/s11229-022-03762-3.

Thomson, Judith Jarvis. 2008. "Turning the Trolley." *Philosophy and Public Affairs* 36 (4): 359–74. https://doi.org/10.1111/j.1088-4963.2008.00144.x.

A The Tipping Line of Linear Model Variants

Linear model variants involve a linear function $G(x) = 1 - x$ in the calculation of account (A), faithfulness (F) and systematicity (S) instead of the quadratic function $G(x) = 1 - x^2$ used in Beisbart, Betz, and Brun (2021). For the linear models, we observed a *tipping line* in ternary plots that marks off configurations of weights that lead to drastically different behaviour with respect to the attainment of full RE states (see, e.g., Figure 5.7) consistency considerations (see, e.g., Figure 6.4 and Figure 6.5), or the maximization of measures such as account or faithfulness (see, e.g., Figure 7.9 and Figure 7.10).

This tipping line is characeterized by the following equation:

$$\alpha_A = \frac{1 - \alpha_S}{2} \tag{A.1}$$

The boundary condition $\alpha_A + \alpha_S + \alpha_F = 1$ allows us to rewrite Equation A.1 in an even simpler form:

$$\alpha_A = \alpha_F$$

Consequently, the tipping line splits the space of weight configurations into the two regions $\alpha_A < \alpha_F$ and $\alpha_A > \alpha_F$.

There are interesting analytical results for both regions. The following propositions and their corollaries help to explain the salient change in the behaviour of linear model variants when crossing the tipping line.[1]

- **Proposition 1**: For the linear model variants all global optima are full RE states if $\alpha_A > \alpha_F$.
- **Proposition 2**: For the linear model variants all global-optimum commitments maximize faithfulness if $\alpha_F > \alpha_A$.

[1] We follow the notation used in Beisbart, Betz, and Brun (2021). We did not explicitly define all terms here. You can find some of the missing definitions in the introduction (Chapter 2) and some in Beisbart, Betz, and Brun (2021).

A.1 Proposition 1

Let τ be a dialectical structure and $\mathcal{C}_0$ some initial commitments. Moreover, assume $\alpha_A > \alpha_F$ for a configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$ in a linear model variant. Then, all global optima (relative to $\mathcal{C}_0$) are full RE states.

Corollaries The linear model variants exhibit the following behaviour for $\alpha_A > \alpha_F$.

- For global optima, there are no inconsistency-preserving cases.
- Consistency-eliminating cases do not occur for global optima.

Proof sketch

Intuitively, $\alpha_A > \alpha_F$ means that account trumps faithfulness. Accordingly, the process can maximize account during the adjustment step of commitments without caring about faithfulness.

Assume that an epistemic state $(\mathcal{C}, \mathcal{T})$ is a global optimum according to the achievement function Z given some initial commitments $\mathcal{C}_0$ and a configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$ such that $\alpha_A > \alpha_F$. We need to show that $(\mathcal{C}, \mathcal{T})$ is a full RE state, i.e., that $\mathcal{T}$ fully and exclusively accounts for $\mathcal{C}$, or equivalently, $A(\mathcal{C}, \mathcal{T}) = 1$.

For a proof by contradiction, assume that

$$A(\mathcal{C}, \mathcal{T}) = G\left(\frac{D_{0,\,0.3,\,1,\,1}(\mathcal{C}, \overline{\mathcal{T}})}{n}\right) < 1$$

This holds only if $D_{0,\,0.3,\,1,\,1}(\mathcal{C}, \overline{\mathcal{T}}) > 0$. In other words, there is at least one sentence s (negated or unnegated) for which there is a positive contribution to the Hamming distance (penalty). In particular, we have the following cases:

1. $\overline{\mathcal{T}}$ extends $\mathcal{C}$ with respect to s: There is $s \in \overline{\mathcal{T}}$, but s and $\neg s$ are not in $\mathcal{C}$.
   - penalty: 0.3
2. $\overline{\mathcal{T}}$ contracts $\mathcal{C}$ with respect to s: There is $s \in \mathcal{C}$, but s and $\neg s$ are not in $\overline{\mathcal{T}}$.
   - penalty: 1
3. $\overline{\mathcal{T}}$ and $\mathcal{C}$ contradict each other with respect to s: Either $s \in \overline{\mathcal{T}}$ and $\neg s \in \mathcal{C}$ or $\neg s \in \overline{\mathcal{T}}$ and $s \in \mathcal{C}$
   - penalty: 1

Each case of changing $\mathcal{C}$ with respect to s, yielding new commitments $\mathcal{C}'$, impacts the contributions to the Hamming distances for account and faithfulness. Note that systematicity is not affected by changing the commitments.

The complete linearity of the achievement function allows us to distribute ("push in") the weights α_A and α_F over the individual contributions of the hamming distances.

$$
\begin{aligned}
&Z(C,T|C_0)\\
&= \alpha_A \cdot A(C,T) + \alpha_F \cdot F(C|C_0) + \alpha_S \cdot S(T)\\
&= \alpha_A \cdot (1 - \frac{D_{0,0.3,1,1}(C,\overline{T})}{n}) + \alpha_F \cdot (1 - \frac{D_{0,0,1,1}(C_0,C)}{n}) + \alpha_S \cdot (1 - \frac{|T|-1}{|\overline{T}|})\\
&= \alpha_A - \frac{\alpha_A \cdot D_{0,0.3,1,1}(C,\overline{T})}{n} + \alpha_F - \frac{\alpha_F \cdot D_{0,0,1,1}(C_0,C)}{n} + \alpha_S - \frac{\alpha_S \cdot (|T|-1)}{|\overline{T}|}\\
&= 1 - \frac{\alpha_A \cdot D_{0,0.3,1,1}(C,\overline{T}) + \alpha_F \cdot D_{0,0,1,1}(C_0,C)}{n} - \frac{\alpha_S \cdot (|T|-1)}{|\overline{T}|}
\end{aligned}
$$

Changing the commitments has no effect on

$$\frac{\alpha_S \cdot (|\mathcal{T}|-1)}{|\overline{\mathcal{T}}|},$$

and n is fixed. Consequently, Z can be optimised by changing the commitments such that the following term is minimized:

$$
\begin{aligned}
&\alpha_A \cdot D_{0,0.3,1,1}(\mathcal{C},\overline{\mathcal{T}}) + \alpha_F \cdot D_{0,0,1,1}(\mathcal{C}_0,\mathcal{C})\\
&= \alpha_A \cdot \sum_{i=1}^{n} d_{0,0.3,1,1}(\mathcal{C},\overline{\mathcal{T}},\{s_i,\neg s_i\}) + \alpha_F \cdot \sum_{i=1}^{n} d_{0,0,1,1}(\mathcal{C}_0,\mathcal{C},\{s_i,\neg s_i\})\\
&= \sum_{i=1}^{n} \alpha_A \cdot d_{0,0.3,1,1}(\mathcal{C},\overline{\mathcal{T}},\{s_i,\neg s_i\}) + \alpha_F \cdot d_{0,0,1,1}(\mathcal{C}_0,\mathcal{C},\{s_i,\neg s_i\})
\end{aligned}
$$

Since the achievement function is optimized for minimal contributions and $\alpha_A > \alpha_F$, it is always more attractive to change the commitments to increase account rather than faithfully respecting the initial commitments. To see this, consider the change in contributions multiplied by the corresponding weights in the table below. This argument can be repeated for every sentence for which $\mathcal{C}$ and $\overline{\mathcal{T}}$ differ.

	account penalty	faithfulness penalty (worst case)
adjusted commitmetments	$d_{0,0.3,1,1}(\mathcal{C}',\overline{\mathcal{T}},\{s,\neg s\})$	$d_{0,0,1,1}(\mathcal{C}_0,\mathcal{C}',\{s,\neg s\})$
-old commitments	$-d_{0,0.3,1,1}(\mathcal{C},\overline{\mathcal{T}},\{s,\neg s\})$	$-d_{0,0,1,1}(\mathcal{C}_0,\mathcal{C},\{s,\neg s\})$

	account penalty	faithfulness penalty (worst case)
change	difference	difference
remove contradicting element from $\mathcal{C}$	-1	+1
revise contradicting element in $\mathcal{C}$	-1	+1
add missing element to $\mathcal{C}$	-0.3	0
remove additional element from $\mathcal{C}$	-1	+1

In summary, if $(\mathcal{C}, \mathcal{T})$ is a global optimum but $A(\mathcal{C}, \mathcal{T}) < 1$, then there is a position $(\mathcal{C}', \mathcal{T})$ such that $A(\mathcal{C}, \mathcal{T}) < A(\mathcal{C}', \mathcal{T})$ contradicting $(\mathcal{C}, \mathcal{T})$ being a global optimum. Consequently, we must have $A(\mathcal{C}, \mathcal{T}) = 1$, i.e., $\mathcal{T}$ accounts fully and exclusively for S. This shows that $(\mathcal{C}, \mathcal{T})$ is a full RE state.

Remark: Note that this argument does not work for quadratic model variants, and in particular, the default model of Beisbart, Betz, and Brun (2021). The Hamming distance D is a summation of penalties. Consequently, squaring the hamming distance yields a polynomial expression where every contributing penalty "interferes" due to multiplication with the others. The resulting multiplicative terms block the above strategy of comparing the contributions and distributing the weights α_A or α_S over these expressions. This is why the quadratic models' share of full RE states among global optima changes gradually with a change in α-weights (see Chapter 5).

A.2 Proposition 2

Assume that a dialectical structure τ and some initial commitments $\mathcal{C}_0$ are given. Moreover, assume $\alpha_A < \alpha_F$ for a configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$ in a linear model variant. Then, for all global optima:

$$F(\mathcal{C} \,|\, \mathcal{C}_0) = 1.$$

Corollaries

The linear model variants exhibit the following behaviour for $\alpha_A < \alpha_F$:

- The relative share of inconsistency-eliminating cases among global optima is 0.0.
  - *Explanation*: Removing or revising an initial inconsistency requires deviating from the initial commitments, which is incompatible with maximal faithfulness.

- Similarly, the relative share of inconsistency-preserving cases in this region of weight configurations corresponds to the relative share of inconsistent initial commitments.
- In turn, the relative share of global optima with maximal value for faithfulness is 1.0.

Proof sketch of Proposition 2

The proof of Proposition 2 is highly similar to that of Proposition 1.

For a proof by contradiction, assume that $(\mathcal{C}, \mathcal{T})$ is a global optimum according to Z, but $F(\mathcal{C} \,|\, \mathcal{C}_0) < 1$.

This holds only if $G\left(\frac{D_{0,0,1,1}(\mathcal{C}_0,\mathcal{C})}{n}\right) < 1$, i.e. only if $D_{0,0,1,1}(\mathcal{C}_0, \mathcal{C}) > 0$. In other words, there is at least one sentence for which there is a positive contribution to the Hamming distance. In particular, there are two cases:

1. $\mathcal{C}$ contracts $\mathcal{C}_0$ with respect to s: +1 (there is $s \in \mathcal{C}_0$, but s and $\neg s$ are not in $\mathcal{C}$)
2. $\mathcal{C}$ and $\mathcal{C}_0$ contradict each other with respect to s: +1

Consider the impacts on individual contributions to the Hamming distances for account and faithfulness of changing $\mathcal{C}$ with respect to s, yielding new commitments $\mathcal{C}'$, in particular the difference $d(\mathcal{C}_0, \mathcal{C}', \{s, \neg s\}) - d(\mathcal{C}_0, \mathcal{C}, \{s, \neg s\})$. In the following subcases, (*) will denote the worst cases.

Case 1

There is an s in $\mathcal{C}_0$, but s and $\neg s$ are not in $\mathcal{C}$. We can now define a new $\mathcal{C}'$ by $\mathcal{C}' := \mathcal{C} \cup \{s\}$

Faithfulness

- agreement (new) - contraction (old): -1

Account

- Case: $s \in \overline{\mathcal{T}}$: agreement (new) - expansion (old): -0.7
- Case $\neg s \in \overline{\mathcal{T}}$: contradiction (new) - expansion (old): + 0.7
- Case s and $\neg s \notin \overline{\mathcal{T}}$ (*): contraction (new)- agreement (old): +1

That is, adding s to $\mathcal{C}$ yields a +1 contribution to the account penalties in the worst case. This is counterbalanced by a -1 improvement in the faithfulness penalties.

Case 2

Without loss of generality, we can assume that $s \in \mathcal{C}_0$ and $\neg s \in \mathcal{C}$. We can now either remove $\neg s$ from $\mathcal{C}$ (Subcase A) or revise $\mathcal{C}$ with s (Subcase B).

Subcase A: $\mathcal{C}' := \mathcal{C} \setminus \{\neg s\}$

Faithfulness

- contraction (new) - contradiction (old): +0

Account

- $s \in \overline{\mathcal{T}}$: expansion (new) - contradiction (old): -0.7
- Case $\neg s \in \overline{\mathcal{T}}$ (*): expansion (new) - agreement (old): +0.3
- Case s and $\neg s \notin \overline{\mathcal{T}}$: agreement(new) - contraction(old): -1

Now, removing $\neg s$ from $\mathcal{C}$ leads to a worsening in the account penalties of +0.3 in the worst case. This is contrasted with no differences in the contributions to faithfulness.

Subcase B: $\mathcal{C}' := (\mathcal{C} \setminus \{\neg s\}) \cup \{s\}$

Faithfulness

- agreement (new) - contradiction (old): -1

Account

- Case: $s \in \overline{\mathcal{T}}$: agreement (new) - contradiction (old): -1
- Case $\neg s \in \overline{\mathcal{T}}$ (*): contradiction (new) - agreement (old): +1
- Case s and $\neg s \notin \overline{\mathcal{T}}$: contraction (new) - contraction (old): +0

In this case, revising $\mathcal{C}$ with s leads to a +1 contribution to the account penalties in the worst case. This is counterbalanced by an improvement of -1 in the faithfulness penalties.

The complete linearity of the achievement function allows us to distribute (push in) the weights α_A and α_F over the individual contributions of the hamming distances in Z. Hence, the weights also apply to the individual contributions considered above. Moreover, changing the commitments does not affect the systematicity of the theory, i.e. $S(\mathcal{T})$ is identical for $(\mathcal{C}, \mathcal{T})$ and $(\mathcal{C}', \mathcal{T})$. Hence, the achievement function is optimized for minimal contributions in the measures for account and faithfulness and $\alpha_F > \alpha_A$.

Consequently, it is always (Case 1, Case 2 (A and B)) more attractive to stay faithful to the initial commitments rather than to change the commitments in order to increase account.

This argument can be repeated for every sentence, for which $\mathcal{C}_0$ and $\mathcal{C}$ differ.

In summary, if $(\mathcal{C}, \mathcal{T})$ is a global optimum but it is assumed that $F(\mathcal{C} \,|\, \mathcal{C}_0) < 1$, then there is a position $(\mathcal{C}', \mathcal{T})$ such that $Z(\mathcal{C}, \mathcal{T} \,|\, \mathcal{C}_0) < Z(\mathcal{C}', \mathcal{T} \,|\, \mathcal{C}_0)$, contradicting $(\mathcal{C}, \mathcal{T})$ being a global optimum. Consequently, we must have $F(\mathcal{C} \,|\, \mathcal{C}_0) = 1$.

A.3 Generalization to Fixed Points

The results we prooved for the linear model variants hold not only for global optima but also for fixed points, which requires but a slight modification of the above proofs. The following proof sketch shows how to generalize Proposition 1 to fixed points for the semi-globally optimizing model variant `LinearGlobalRE`. Proposition 2 can be generalized similarly.

Proof sketch

Let τ be a dialectical structure and $\mathcal{C}_0$ some initial commitments. Moreover, assume $\alpha_A > \alpha_F$ for a configuration of weights $(\alpha_A, \alpha_S, \alpha_F)$.

For a proof by contradiction, we assume that $(\mathcal{C}_i, \mathcal{T}_i)$ is a fixed point with $A(\mathcal{C}_i, \mathcal{T}_i) < 1$.

$(\mathcal{C}_i, \mathcal{T}_i)$ being a fixed point implies that $(\mathcal{C}_{i-1}, \mathcal{T}_{i-1}) = (\mathcal{C}_i, \mathcal{T}_i)$ and hence that $A(\mathcal{C}_{i-1}, \mathcal{T}_{i-1}) < 1$ as well. However, during the last adjustment step (from $i-1$ to i), *all* minimally consistent positions were available as candidates. Since $A(\mathcal{C}_{i-1}, \mathcal{T}_{i-1}) < 1$, the process could have found *other* commitments $\mathcal{C}_i'$ which would have resulted from changing $\mathcal{C}_{i-1}$ with respect to s following the same line of reasoning we used to prove Proposition 1. Again, there would have been at least one sentence s for which there is a positive contribution to the Hamming distance in the measure of account. Hence, there would have been $(\mathcal{C}_i', \mathcal{T}_i)$ with $\mathcal{C}_i' \neq \mathcal{C}_{i-1}$ that would have performed better than $(\mathcal{C}_i, \mathcal{T}_i)$ according to the achievement functon. This shows that $(\mathcal{C}_i, \mathcal{T}_i)$ cannot be a fixed point (contradicting the assumption).

Local Model variants

Finally, we can also generalize Proposition 1 and Proposition 2 to fixed points of the `LinearLocalRE` model variant. The difference to the semi-globally optimizing RE process of `LinearGlobalRE` is that locally optimizing models (with a neighborhood depth of 1) proceed by changing at most one sentence per adjustment step. But this is all we need to get the above proofs by contradiction off the ground, where we only considered hypothetical adjustments of the commitments with respect to a single sentence. Accordingly, the propositions will also hold if we enlarge the d-neighborhood to more than one sentence.

B Trivial Endpoints

B.1 Background

A "trivial" endpoint is a fixed point or a global optimum that consists of a singleton theory (e.g. $T = \{s_1\}$) and a singleton commitment (e.g. $C = \{s_1\}$).

Such outcomes are not bad per se, but they may be indicative of the model exploiting shortcomings in the underlying measures. In particular, "trivial" endpoints may be a consequence of the original model's shortcoming concerning the measure of systematicity, which does not discriminate between singleton theories on the basis of the scope of theories. Note that the same shortcoming also applies to the model variants explored in this report.

B.2 Results

> **Note**
>
> The results of this chapter can be reproduced with the Jupyter notebook located here.

B.2.1 Overall Results

Model	Relative share of trivial global optima	Number of trivial global optima	Number of global optima
QuadraticGlobalRE	0.009	6625	714584
LinearGlobalRE	0.081	56635	700830
QuadraticLocalRE	0.009	6625	709289
LinearLocalRE	0.07	50256	721096

Table B.1: Relative share of trivial global optima

Model	Relative share of trivial fixed points	Number of trivial fixed points	Number of fixed points
QuadraticGlobalRE	0.008	3698	458147
LinearGlobalRE	0.08	25111	312783
QuadraticLocalRE	0.009	5189	588236
LinearLocalRE	0.063	14443	228122

Table B.2: Relative share of trivial fixed points (result perspective)

Model	Relative share of trivial fixed points	Number of trivial fixed points	Number of fixed points
QuadraticGlobalRE	0.007	3700	528616
LinearGlobalRE	0.08	25111	313002
QuadraticLocalRE	0.006	11652	1991852
LinearLocalRE	0.323	421058	1303077

Table B.3: Relative share of trivial fixed points (process perspective)

Observations

- Overall, the relative share of trivial gobal optima (Table B.1) and fixed points (result perspective Table B.2) is very low for quadratic model variants
- Linear model variants exhibit substantially more trivial global optima, but the relative shares are still low.
- `LinearLocalRE` exhibits a substantial share of trivial fixed points in the process perspective (Table B.3), but not for the result perspectve (Table B.2). This indicates that relatively many branches lead to trivial fixed points.

B.2.2 Results Grouped by Sentence Pool Size

Observations

- The relative shares of trivial global optima or fixed points tend to decrease with increasing sentence pool sizes.
- A notable exception to this trend is `LinearLocalRE` in the process perspective (Figure B.3)

B.2.3 Results Grouped by Configuration of Weights

Observations

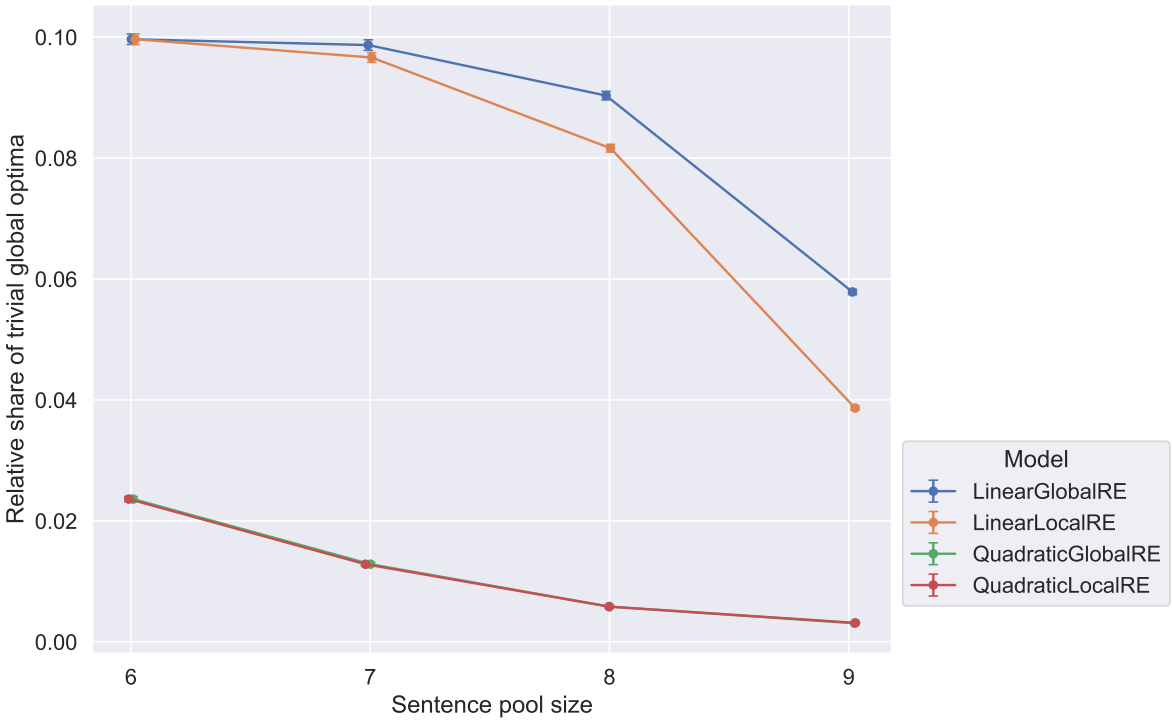


Figure B.1: Relative share of trivial global optima grouped by model variant and sentence pool size

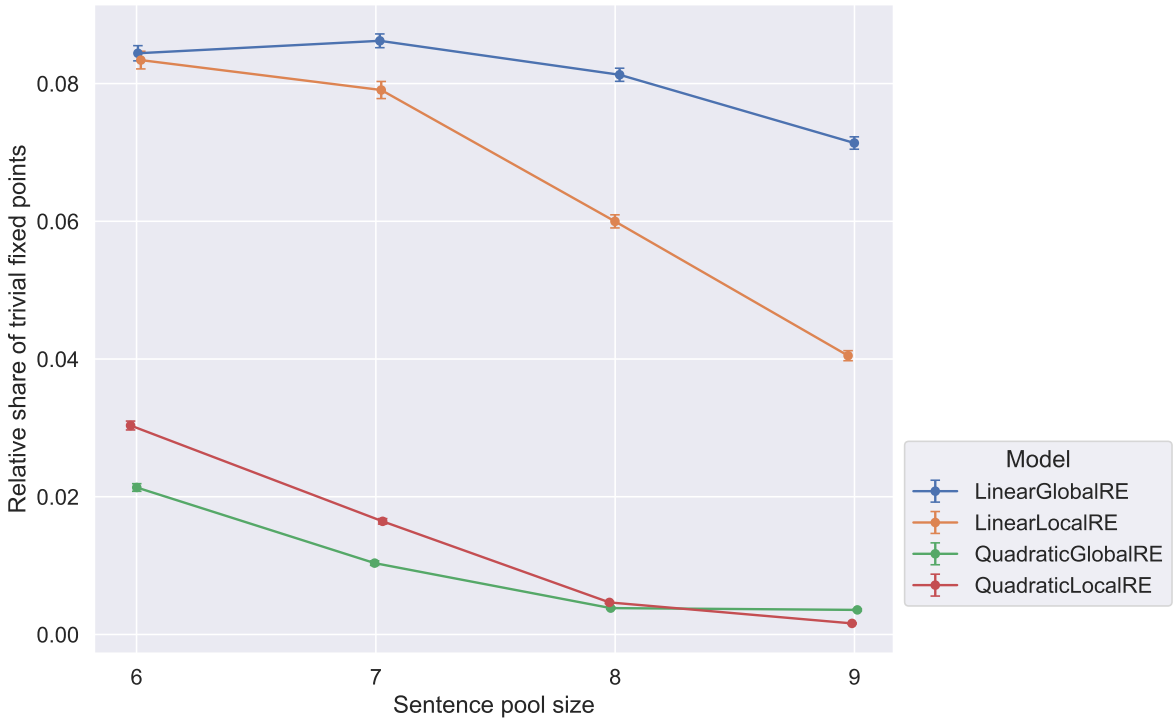


Figure B.2: Relative share of trivial fixed points (result perspective) grouped by model variant and sentence pool size

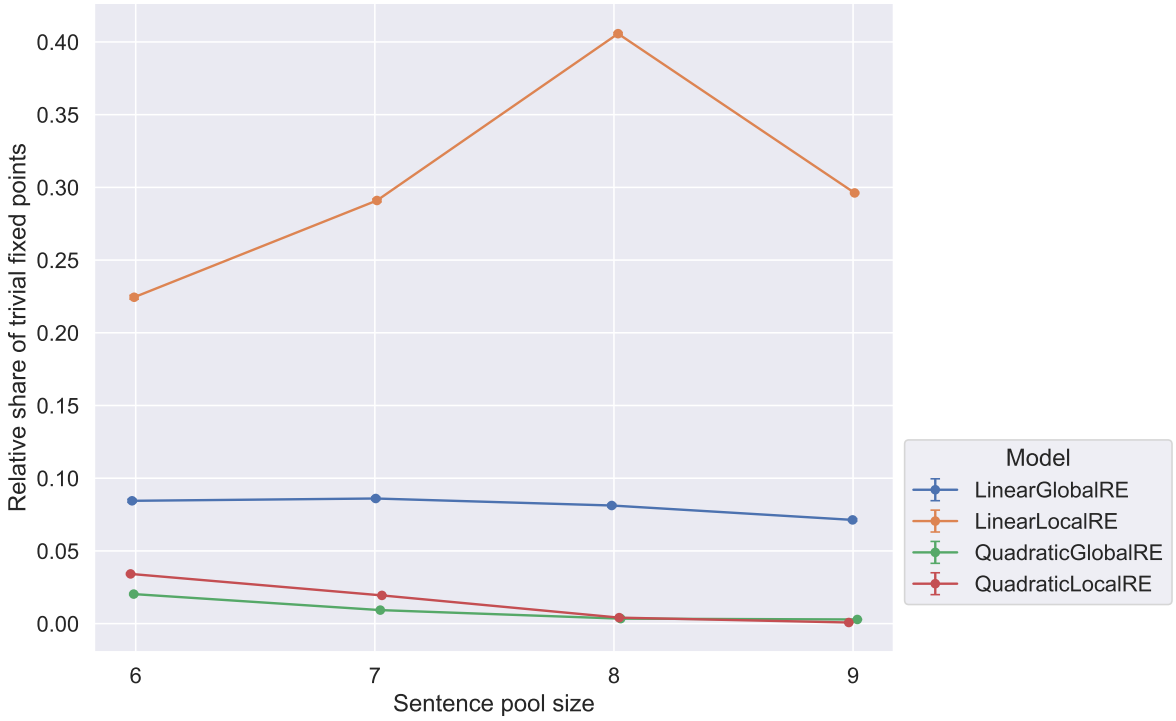


Figure B.3: Relative share of trivial fixed points (process perspective) grouped by model variant and sentence pool size

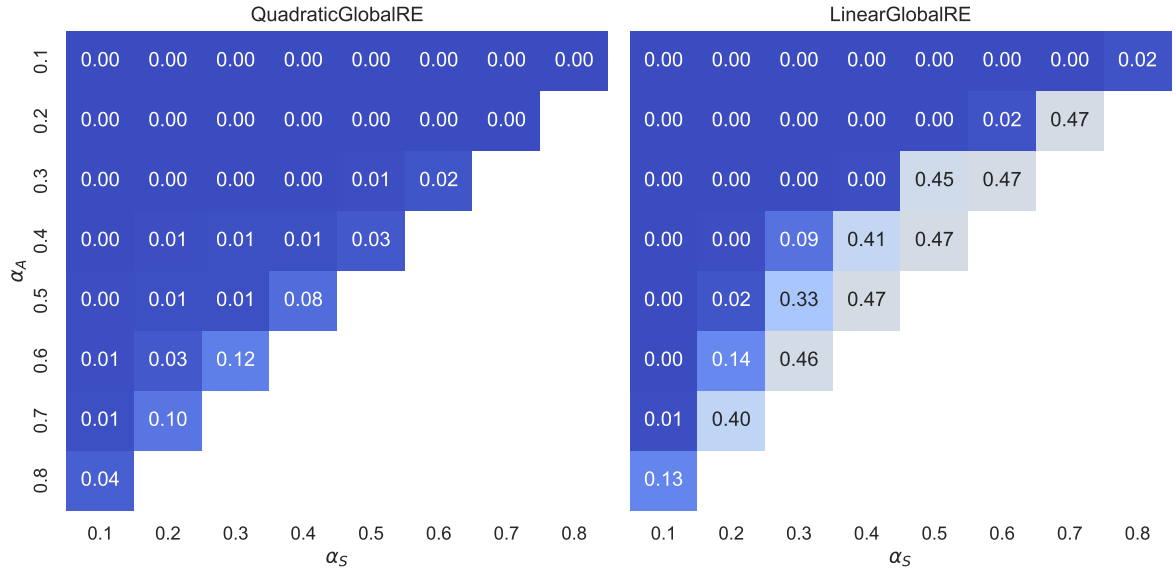


Figure B.4: Relative share of trivial global optima grouped by model variant and weight configuration

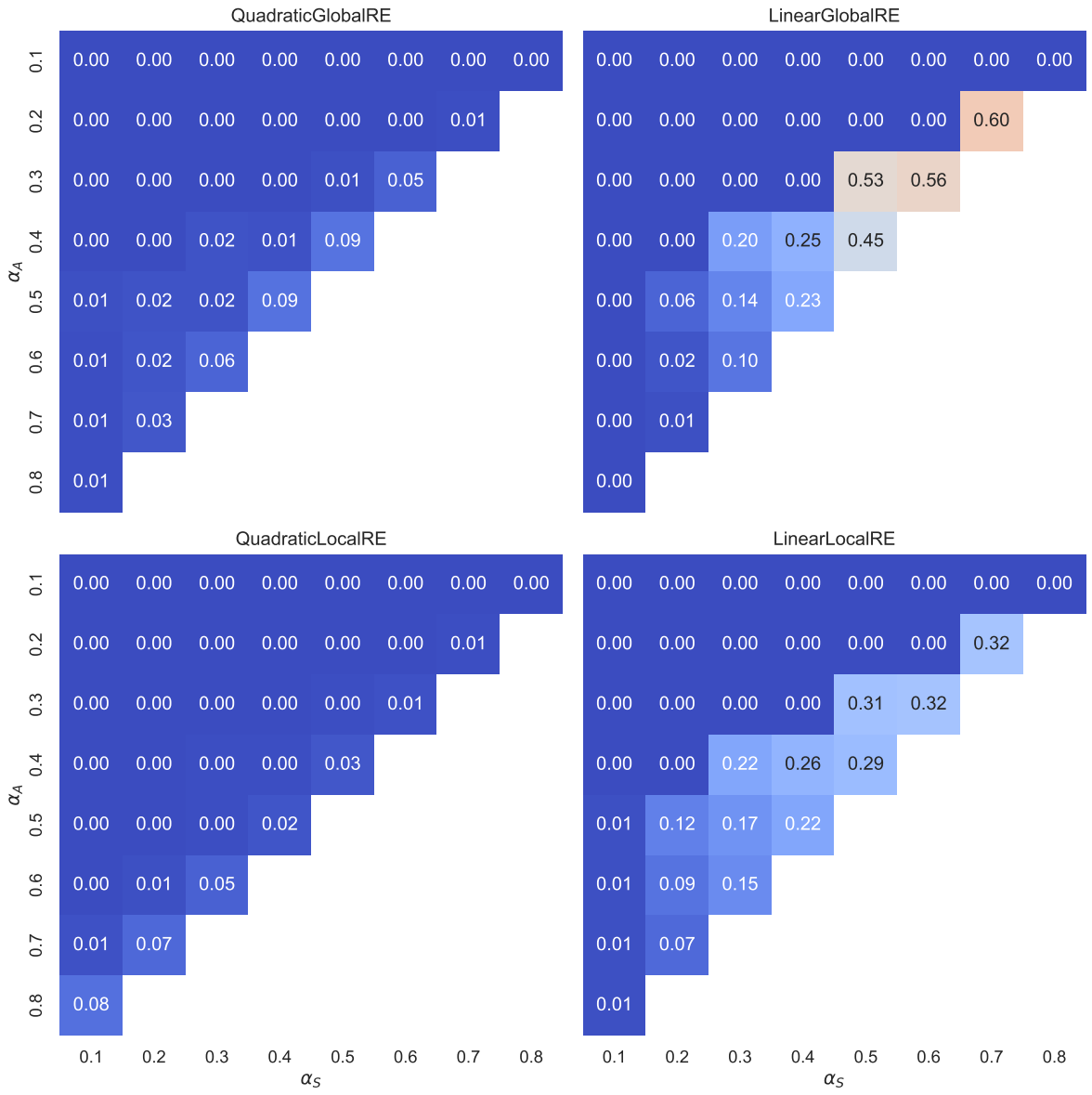


Figure B.5: Relative share of trivial fixed points (result perspective) grouped by model variant and weight configuration

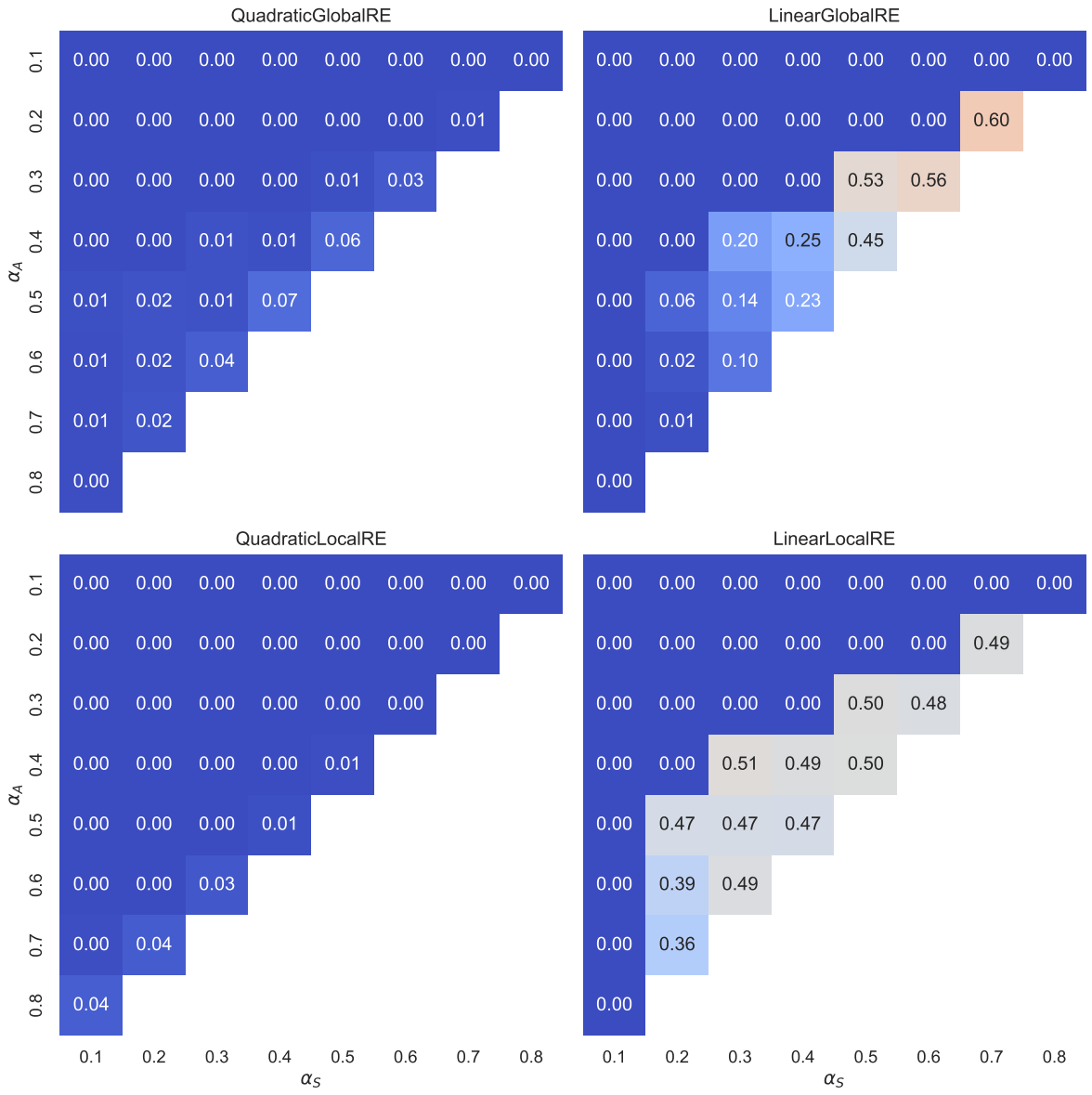


Figure B.6: Relative share of trivial fixed points (process perspective) grouped by model variant and weight configuration

- In quadratic model variants, the configuration of weights have a small impact on the relative shares of trivial endpoints.
- Linear model variants tend to produce higher relative shares of trivial endpoints for low values of α_F and high values of α_S.

C Alternative Systematicity Measures

In this appendix, we will point to some shortcomings of the systematicity measure used in Beisbart, Betz, and Brun (2021) and discuss several alternative measures.[1]

> **Note**
>
> The results of this appendix can be reproduced with the following Jupyter notebook: https://github.com/re-models/re-technical-report/blob/main/notebooks/appendix_systematicity_measures.ipynb.

C.1 Desiderata for systematicity measures

C.1.1 D1 – Content

The achievement function models the trade-off between the three desiderata *account*, *faithfulness* and *systematicity*. The latter is supposed to measure the extent of a theory's ability to systematize sentences from the given sentence pool $\mathcal{S}$. The formulation is admittedly in need of explication. Beisbart, Betz, and Brun (2021) used the following definition for their RE model:

$$S_{BBB}(\mathcal{T}) = 1 - \left(\frac{|\mathcal{T}| - 1}{|\overline{\mathcal{T}}|} \right)^2 \tag{C.1}$$

with $\mathcal{T}$ being a set of sentences representing the principles of the theory and $\overline{\mathcal{T}}$ being the dialectical closure of $\mathcal{T}$ (i.e., all implications of $\mathcal{T}$ according to some dialectical structure τ).

The underlying idea is simple. The more content a theory has (as, for instance, measured by the amount of its implications), the more sentences it systematizes. Hence, we should require:

Content (D1): Everything else being equal, systematicity should (monotonically) increase with increasing content.

[1] The considerations and suggestions we present in this appendix are based on different project-internal drafts and were discussed and further developed on several occasions within our project group of the project 'How far does Reflective Equilibrium Take us? Investigating the Power of a Philosophical Method'. The considerations presented here are, in particular, not our original ideas.

C.1.2 D2 – Simplicity

This simple suggestion is, however, in need of refinement. The systematizing power of a theory should be evaluated in relation to its size. If a theory implies many sentences only because it contains many sentences as its principles, its systematizing power should be considered low. The reason is that systematization is usually thought of as somehow summarising a lot with little. Theories in physics systematize empirical facts to the extent that they imply a lot of these facts by using but few physical laws (as, for instance, Newton's three laws of motion).

These considerations motivate:

Simplicity (D2): Everything else being equal, systematicity should (monotonically) increase with decreasing theory size.

How does the suggested measure S_{BBB} conform to these constraints? In our modelling context, a theory is simply a set of sentences, which you can think of as its principles or some axiomatic basis. Accordingly, the size of a theory can be measured by $|\mathcal{T}|$ in Equation C.1. There are different possibilities for conceptualizing the notion of content. One suggestion is to equate the dialectical closure of a theory ($\overline{\mathcal{T}}$) with its content.

Since the sentence pool is finite, so is the dialectical closure of a theory.[2] Accordingly, we can measure the size of a theory's content by $|\overline{\mathcal{T}}|$.

The bracketed term in Equation C.1 can be considered as a penalizing contribution, which increases with the theory's size ($|\mathcal{T}|$) and decreases with the theory's content size ($|\overline{\mathcal{T}}|$). Figure C.1 illustrates systematicity values calculated by S_{BBB} for a sentence pool of size 14.[3]

By following vertical lines (constant theory closure size), you can see that everything else being equal, "smaller" theories receive higher systematicity values. Hence, S_{BBB} satisfies *D2 (simplicity)*. By following the plotted lines (constant theory size), you can see that S_{BBB} satisfies *D1 (content)* for all theory sizes except for $|\mathcal{T}| = 1$. As noted before (see Chapter 7), these singleton theories receive the maximal systematicity value of 1.0 independent of their content.

How problematic is this violation of *D1*? After all, *D1* is only violated for singleton theories and only violated in a "weak" sense. While it is true that systematicity does not *monotonically* increase with increasing content for singleton theories, systematicity does at least not decrease with increasing content. In Chapter 7, we observed that fixed points and global optima frequently maximize the standard measure of systematicity (with singleton theories). In Appendix B, we presented a preliminary analysis of how pervasive fixed points and global optima are that consist of a singleton theory and a single commitment. The sparse emergence

[2] The closure $\overline{\mathcal{T}}$ is defined by $\{s \in \mathcal{S} | \mathcal{T} \models_\tau s\}$, where $\models_\tau$ denotes the relation of dialectical implication. In other words, this closure only contains "atomic" sentences from the sentence pool; it does not include any other logical consequences (such as conjunctions, for example).

[3] In the RE Model, theories are dialectically consistent and therefore minimally consistent (see Chapter 2). Hence, a theory can have at most n principles if $2n$ is the size of the sentence pool.

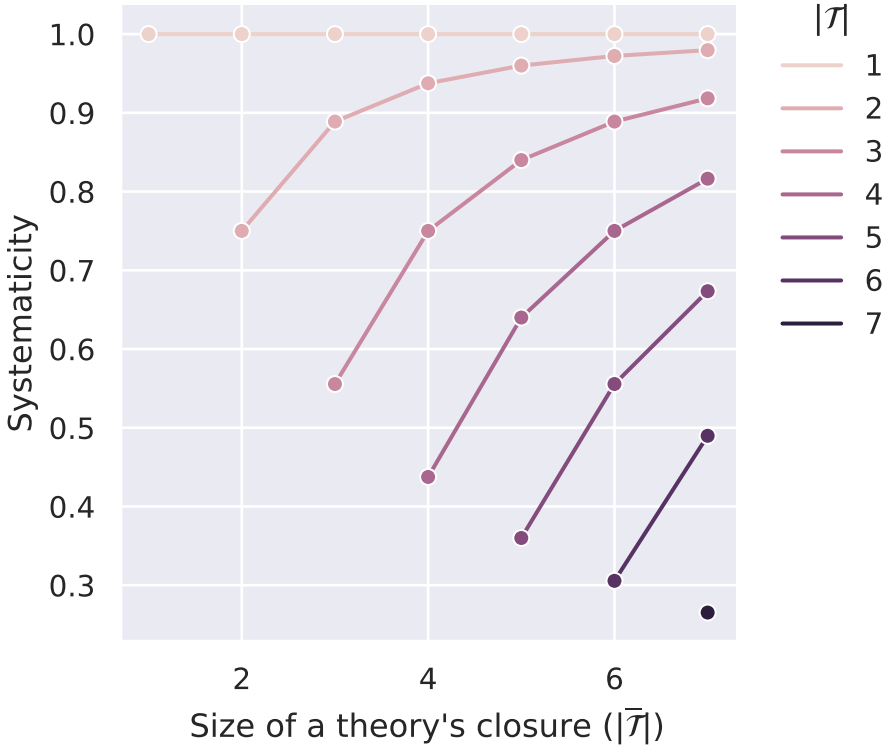


Figure C.1: Standard systematicity of theories in dependence of their size and closure's size.

of such "trivial" endpoints suggests that singleton theories (with extremely low content) do not have a significant advantage over other theories. But this does not mean that the violation of *D1* could not lead to problematic behavior of the model in other contexts. We should, therefore, consider and analyze other systematicity measures, which we intend in this appendix.

C.1.3 D3 – Minimal Systematicity

There are other constraints as well: Due to the assumption that the sentence pool is finite, there are lower bounds and upper bounds for systematicity. The systematicity measure S_{BBB} is normalized to yield values within the unit interval $[0, 1]$. We will follow this convention.

So, under which conditions should systematicity be minimal? The above formulated intuitions that led to *D1* and *D2* suggest that

Minimal systematicity (D3): Systematicity should be minimal if a theory does not imply anything besides its principles.

We might say that theories that do not imply anything in addition to their principles are vacuous in the sense of being ineffective in their aim to systematize sentences. We will call such theories *ineffective theories*. Similarly, we will call theories that imply more than their principles *effective theories*.

The formulation *D3* is imprecise or even ambiguous. If we read it strongly, we might require:

Minimal systematicity (D3.1): Theories that do not imply anything besides their principles (ineffective theories) receive lower systematicity values than other theories.

In other words, the systematicity values of ineffective theories are lower bounds for effective theories. One way of satisfying *D3.1* is to let $S(\mathcal{T}) = 0$ if $\mathcal{T}$ is an ineffective theory. But there are other possibilities. In particular, *D3.1* allows it to distribute different systematicity values to ineffective theories.

The measure S_{BBB} does not satisfy *D3.1*—not only because of its preferential treatment of singleton theories. In the following, we will refer to points in figures such as Figure C.1 by using tuples of the form $(|\mathcal{T}|, |\overline{\mathcal{T}}|)$. For instance, the point $(3, 4)$ denotes the equivalence class of theories of size three with a dialectical closure of size four. Ineffective theories are points of the form (n, n), which are the left (lower) end points of lines in Figure C.1. You can see in this figure that for $2n = 14$, there are only four theories that are lower bounds for effective theories (namely, $(4, 4)$, $(5, 5)$, $(6, 6)$, $(7, 7)$). For the other ineffective theories, we can find effective theories that receive lower systematicity values (e.g., $S_{BBB}(3, 3) > S_{BBB}(6, 7)$). Hence, S_{BBB} violates *D3.1*.

There is, however, a weaker interpretation of *D3*. We might only demand that ineffective theories receive the lowest systematicity value in comparison to effective theories with the same amount of principles (e.g., $S(3, 3) < S(3, 4) < \cdots < S(3, n)$). This weaker criterion is satisfied

by S_{BBB}. Since this weak version of *D3* is already implied by *D1 (content)*, we will not list it as an additional criterion.

If ineffective theories are, in some sense, the least systematizing, we might ask which theories are most systematizing. According to the above-formulated intuitions, we might suggest that theories with the least number of principles and the largest number of implications should receive maximum systematicity values. For a sentence pool of size 14, these are singleton theories that imply seven sentences. Similar to the weak version of *D3*, this criterion is satisfied by S_{BBB} and already implied by *D1* and *D2*.

C.1.4 D4 – Non-Ad-Hocness

Are there other reasonable constraints we should put on systematicity measures? Consider a theory $\mathcal{T}$ with one sentence ($\mathcal{T} = \{s_1\}$) that has an additional sentence s_2 in its closure ($\overline{\mathcal{T}} = \{s_1, s_2\}$). Suppose further we add another sentence s_3 to construct a new theory $\mathcal{T}^* = \{s_1, s_3\}$. If, now, the dialectical closure $\overline{\mathcal{T}}^*$ is not expanded as compared to $\overline{\mathcal{T}}$ besides the added sentence (i.e., $\overline{\mathcal{T}^*} = \{s_1, s_2, s_3\}$), we will say that we constructed $\mathcal{T}^*$ by adding *ad hoc principles* to $\mathcal{T}$.

One could argue that adding ad hoc principles should not lead to an increase in systematicity.

First, while the dialectical closure does increase by one sentence, the size of the theory is also increased by one. What we win in content, we lose in simplicity. In other words, the introductory intuitions that led to *D1 (content)* and *D2 (simplicity)* might be used to argue that adding ad hoc principles should not increase its systematicity.

Second, there is another intuition we have not used so far. Usually, we think of a theory's principles as working together in their sytematizing activity. For many, or at least for some implications, we have to combine principles. By definition, ad hoc principles do not work together with other principles to imply other sentences. Accordingly, they do not add something to the systematization efforts of the other principles. They work on their own.

Hence, we should require:

D4 (non-ad-hocness): Extending a theory with ad hoc principles (i.e., principles that do not expand the theory's content besides the added principles) should not increase its systematicity.

In the context of modelling RE, *D4* is even too weak to allow the model to penalize the addition of ad hoc principle in every case (independent of the chosen weights). Suppose two theories $\mathcal{T}$ and $\mathcal{T}^*$ where the latter is constructed by adding an ad hoc principle to the former. Suppose further a set of commitments that coincide with the closure of $\mathcal{T}^*$. Additionally, we assume that there are no other theories that compare better with respect to the summation of account and systematicity. In such cases, the achievement function will always prefer $\mathcal{T}^*$ over $\mathcal{T}$ if we don't strengthen *D4*. The problem is that $\mathcal{T}^*$ performs better than $\mathcal{T}$ with respect to account since account is maximized if the theory's closure matches the commitments. We must, therefore,

counterbalance the advantage in account of $\mathcal{T}^*$ over $\mathcal{T}$ by penalizing the addition of ad hoc principles within the systematicity measure.[4] This might suggest that the extension of ad hoc principles should decrease a theory's systematicity.

However, that might be too strong since one might want to satisfy *D3.1* by letting $S = 0$ for ineffective theories. But then, one cannot further reduce systematicity for ad hoc extensions of ineffective theories. Hence, *D.3* might conflict with the requirement that systematicity should decrease with ad hoc extension. Fortunately, there is a simple solution. The described case is only relevant for effective theories. Hence, an appropriate strengthening of *D4* is:

D4.1 (non-ad-hocness): Extending a effective theory with ad hoc principles should (monotonically) decrease its systematicity; extending an ineffective theory with ad hoc principles must not increase its systematicity.

Figure C.1 illustrates that S_{BBB} complies with *D4.1*. This requirement is satisfied if $S(n, m) > S(n + i, m + i)$ (with n the theorie's size, m its closure's size and i the amount of added ad hoc principles). In Figure C.1, you see, for instance, $S_{BBB}(1, 4) > S_{BBB}(2, 5) > S_{BBB}(3, 6) > S_{BBB}(4, 7)$.

C.1.5 D5 – Internal Connectedness

One rationale for *D4 (non-ad-hocness)* was the intuition that ad hoc principles are loners in some way. They do not work together with other principles in implying other sentences than the theory's principles; they do not add something to the inferential potential of a theory besides themselves. The requirement *D4* is, therefore, a special case of a more general requirement that demands:

D5 (internal connectedness): Everything else being equal (content and size), a theory in which principles work together is more systematic than a theory in which principles do not work together (so much).

At this point, we do not further explicate the notion of working together but simply offer two illustrating examples.

Example C.1 (First example for *D5*)**.** Consider the dialectical structure depicted in Figure C.2 and the theories $\mathcal{T}_1 = \{1, 2\}$ and $\mathcal{T}_2 = \{7, 8\}$. Both theories have the same size ($|\mathcal{T}_1| = |\mathcal{T}_2| = 2$) and the same size of their dialectical closure ($|\overline{\mathcal{T}}_1| = |\overline{\mathcal{T}}_2| = 6$). The principles of $\mathcal{T}_1$ work together in the following sense: We need both principles to deduce the other sentences of its dialectical closure ($\{3, 4, 5, 6\}$). In contrast, the principles of $\mathcal{T}_2$ do not work together. Instead, the inferential workload is distributed among its principles: The principle 7 implies 3 and 4, and principle 8 implies 5 and 6. According to *D5*, we should expect that $S(\mathcal{T}_2) < S(\mathcal{T}_1)$

[4] There, of course, other possibilities to adapt the model to achieve this goal.

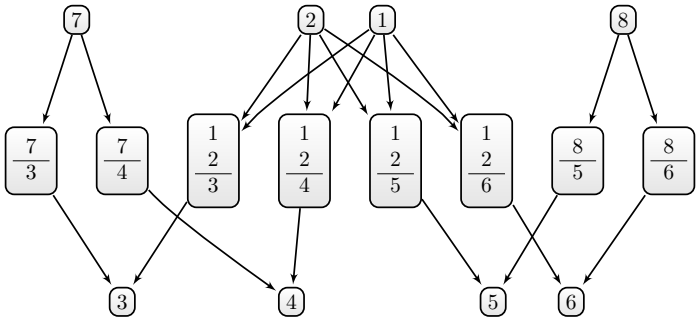


Figure C.2: First illustration of principles (not) working together.

Example C.2 (Second example for *D5*)**.** A similar case is depicted in Figure C.3. Here, you do not need all principles of the theory $\mathcal{T}_1 = \{1, 2, 4\}$ to deduce sentence 3 or 5. However, the principles of $\mathcal{T}_1$ still work together since you need sentence 1 in either case. In contrast, the principles of the theory $\mathcal{T}_2 = \{1, 6, 7\}$ work alone to deduce 3 and 5. (Here, 1 is even an ad hoc principle.) Again, *D5* requires that $S(\mathcal{T}_2) < S(\mathcal{T}_1)$

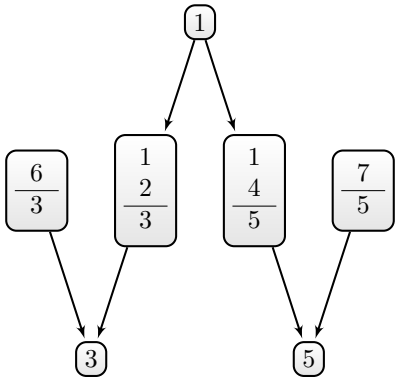


Figure C.3: Second illustration of principles (not) working together.

The measure S_{BBB} cannot satisfy *D5* for the simple reason that the measure is blind to the differences in the given examples. This measure calculates systematicity based on the theory's size and the size of its dialectical closure without considering any other inferential properties of the dialectical structure.

C.1.6 D6 – External Connectedness

So far, we have only considered the inferential potential of a theory based on what is implied by the principles alone (*D1*) and how the principles work together in producing their content

(*D5*). It might, additionally, be relevant to consider what the theory implies with the help of other sentences.

Example C.3 (Example for *D6*)**.** For instance, the theory $\mathcal{T}_1 = \{1\}$ does not imply anything on its own (besides its principle) and is thus on par with other singleton theories according to the original measure of systematicity.However, in contrast to, let's say, the theory $\mathcal{T}_2 = \{4\}$, $\mathcal{T}_1$ does imply sentences if it is combined with other sentences, in particular 2 or 3. We might, therefore, expect that the systematicity of $\mathcal{T}_1$ is higher than the systematicity of $\mathcal{T}_2$.

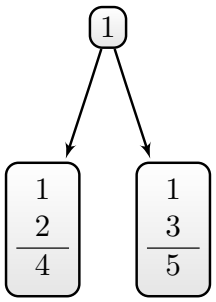


Figure C.4: Illustration of principles working together with other sentences.

This motivates:

D6 (external connectedness): Everything else being equal, if the content of a theory $\mathcal{T}_1$ is larger with some auxiliary assumptions as compared to another theory $\mathcal{T}_2$, then $\mathcal{T}_1$ has a larger systematicity than $\mathcal{T}_2$.

Again, S_{BBB} cannot satisfy *D6* since it is confined to calculate systematicity based on $|\mathcal{T}|$ and $|\overline{\mathcal{T}}|$.

C.2 Simple Systematicity Measures

The measure S_{BBB} uses only the size of a theory and the size of its dialectical closure to calculate systematicity. We will call systematicity measures that follow this recipe *simple systematicity measures*. In the following, we will suggest alternative systematicity measures and analyze their performance concerning *D1-D6*. We will begin with simple systematicity measures.

C.2.1 Minimal Mutation Systematicity

The measure S_{BBB} violates *D1 (content)* due to the numerator $|\mathcal{T}| - 1$ in Equation C.1, which becomes zero for singleton theories. Accordingly, singleton theories receive maximum

systematicity independent of their content. One simple suggestion to fix this behaviour is to adapt the numerator such that it does not become zero for theories of size one. A minimal adaption would be to subtract smaller values than one:

$$S_{mm}(\mathcal{T}|\gamma) := G\left(\frac{|\mathcal{T}| - \gamma}{|\overline{\mathcal{T}}|}\right)$$

with $\gamma < 1$.

Figure C.5 plots the systematicity measure for different values of the parameter γ. By construction, the measure satisfies *D1 (content)*. Similar to S_{BBB}, it also satisfies *D2 (simplicity)* and *D4.1 (non-ad-hocness)*. It is even possible to comply with *D3.1 (minimal systematicity)* if we set γ high enough. In our case (sentence pool of size *14*), $\gamma = 0.1$ is able to push the systematicity values of (m, m) theories (i.e., ineffective theories) such that they are lower bounds for effective theories.

C.2.2 Effective Content Systematicity

The basic idea of the measure S_{BBB} to satisfy *D1 (content)* and *D2 (simplicity)* is to employ the "penalizing" term $\frac{|\mathcal{T}|-1}{|\overline{\mathcal{T}}|}$, which gets bigger with an increase in theory size ($|\mathcal{T}|$) and a decrease in the size of the closure ($|\overline{\mathcal{T}}|$). There are, however, other ideas to implement a similar behaviour. A straightforward suggestion is to use the non-trivial content—that is, a theory's dialectical implications besides its principles ($\overline{\mathcal{T}} \setminus \mathcal{T}$)—to measure systematicity. In this way, an increase in the amount of principles leads to a decrease in systematicity and an increase in the content to an increase.

What remains is a proper normalization of the measure:

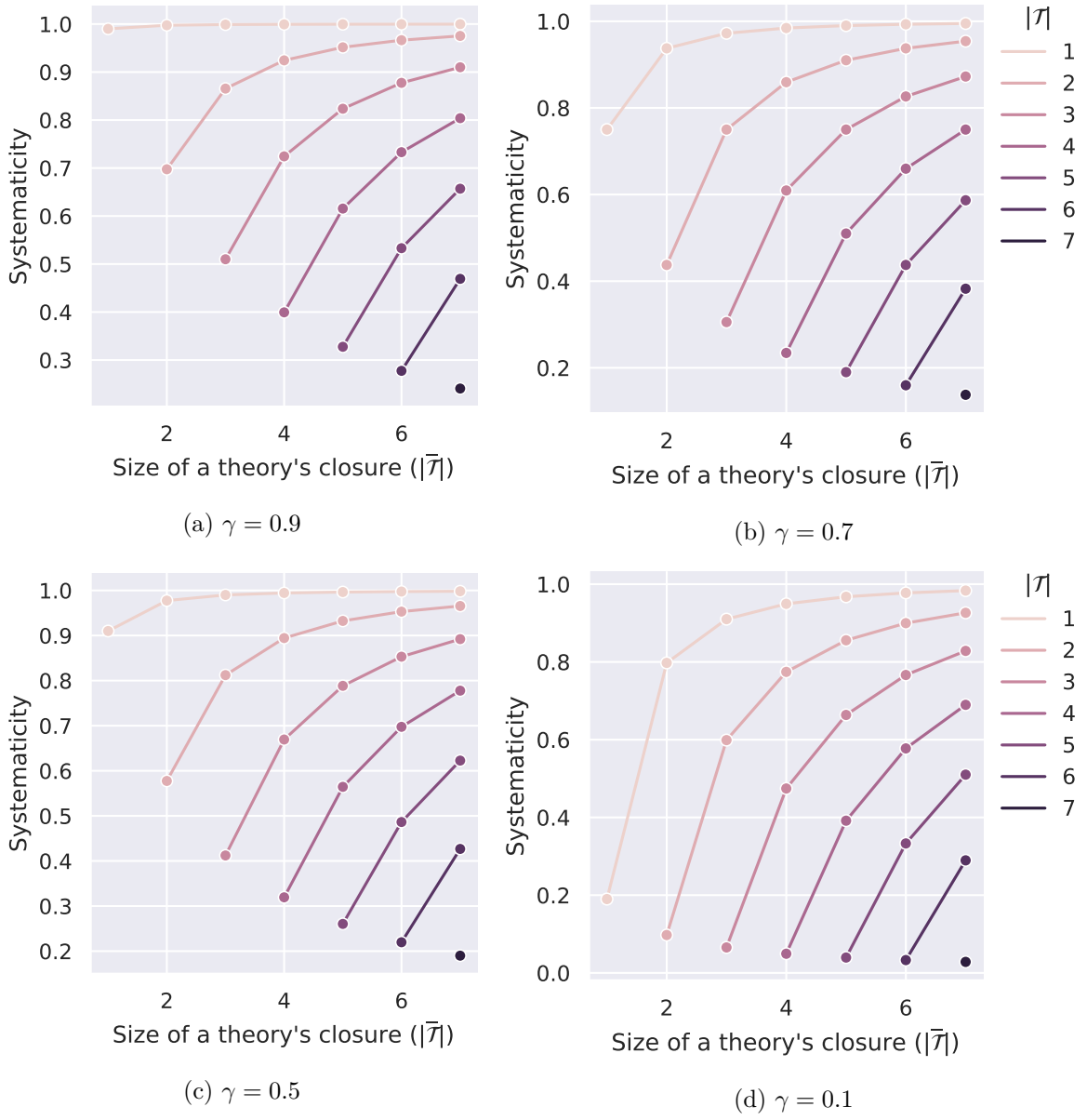
$$S_{ec}(\mathcal{T}) = \frac{|\overline{\mathcal{T}} \setminus \mathcal{T}|}{n-1} = \frac{|\overline{\mathcal{T}}| - |\mathcal{T}|}{n-1}$$

Maximally systematizing theories are singleton theories that are able to imply for every sentence s outside of their *domain* either s or its negation.[5] For such theories, we have $|\overline{\mathcal{T}}| - |\mathcal{T}| = n - 1$, which motivates the denominator. Worst cases are ineffective theories for which $|\overline{\mathcal{T}}| = |\mathcal{T}|$ holds, which yields $S_{ec} = 0$.

S_{ec} is linear. An alternative would be to use a quadratic term that is more akin to the quadratic form of S_{BBB}:

$$S_{ec^2}(\mathcal{T}) = 1 - \left(1 - \frac{|\overline{\mathcal{T}}| - |\mathcal{T}|}{n-1}\right)^2$$

[5] If $\mathcal{S}$ is the sentence pool, the *domain* of a theory $\mathcal{T}$ is defined by $\{s \in \mathcal{S} | s \in \mathcal{T} \text{ or } \neg s \in \mathcal{T}\}$.



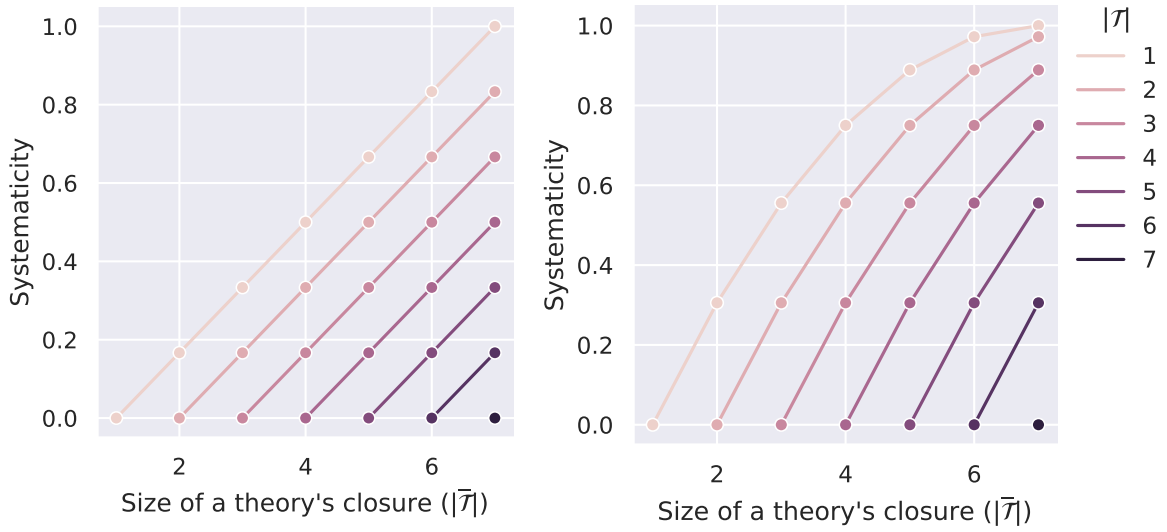
(a) $\gamma = 0.9$

(b) $\gamma = 0.7$

(c) $\gamma = 0.5$

(d) $\gamma = 0.1$

Figure C.5: Minimal mutation systematicity of theories in dependence of their size and closure's size.



(a) Linear effective content systematicity

(b) Quadratic effective content systematicity

Figure C.6: Effective content systematicity of theories in dependence of their size and closure's size.

Both measures satisfy *D1, D2*, and *D3.1*. However, they fail to account for ad hoc principles (*D4.1*). Like all simple measures, they also do not satisfy *D5* and *D6*.

C.2.3 Content-Simplicity Weighted Systematicity

The measure S_{ec} can be motivated in an additional way, which will not only explain why it violates *D4.1 (non-ad-hocness)* but which will allow us to construct other measures which will satisfy *D4.1*.

The basic idea is to formulate separate penalizing terms for simplicity and content:

- Simplicity penalties: $|\mathcal{T}| - 1$
- Content penalties: $n - |\overline{\mathcal{T}}|$

Note that theories that are optimal according to simplicity and content receive no penalties.

We can now aggregate them and introduce an additional parameter α that can be used to balance the penalizing contributions:

$$\alpha \cdot (|\mathcal{T}| - 1) + (1 - \alpha) \cdot (n - |\overline{\mathcal{T}}|) \tag{C.2}$$

Thus, if $\alpha > \frac{1}{2}$, then a loss in simplicity is penalized more severely than a loss in content.

Using $|\overline{\mathcal{T}}| \geq |\mathcal{T}| \geq 0$, one can show that

$$\alpha \cdot (|\mathcal{T}| - 1) + (1 - \alpha) \cdot (n - |\overline{\mathcal{T}}|) \leq |\overline{\mathcal{T}}| \cdot (2 \cdot \alpha - 1) + n \cdot (1 - \alpha) - \alpha$$

Accordingly, we define

$$c := |\overline{\mathcal{T}}| \cdot (2 \cdot \alpha - 1) + n \cdot (1 - \alpha) - \alpha$$

and use it to normalize the penalizing term. We will define the new weighted measure by:

$$S_{csw_\alpha}(\mathcal{T}|\alpha) = 1 - \frac{\alpha \cdot (|\mathcal{T}| - 1) + (1 - \alpha) \cdot (n - |\overline{\mathcal{T}}|)}{c}$$

One can show that $S_{csw_\alpha}(\mathcal{T}|0.5) = S_{ec}$. In other words, if we balance the penalizing terms for content and size similarly, the new measure S_{csw_α} reduces to S_{ec}, which explains why the latter is not able to satisfy *D4.1 (non-ad-hocness)*. Adding an ad hoc principle to a theory will increase its size by one and similarly increase its content by one. What is gained in content is lost in simplicity.

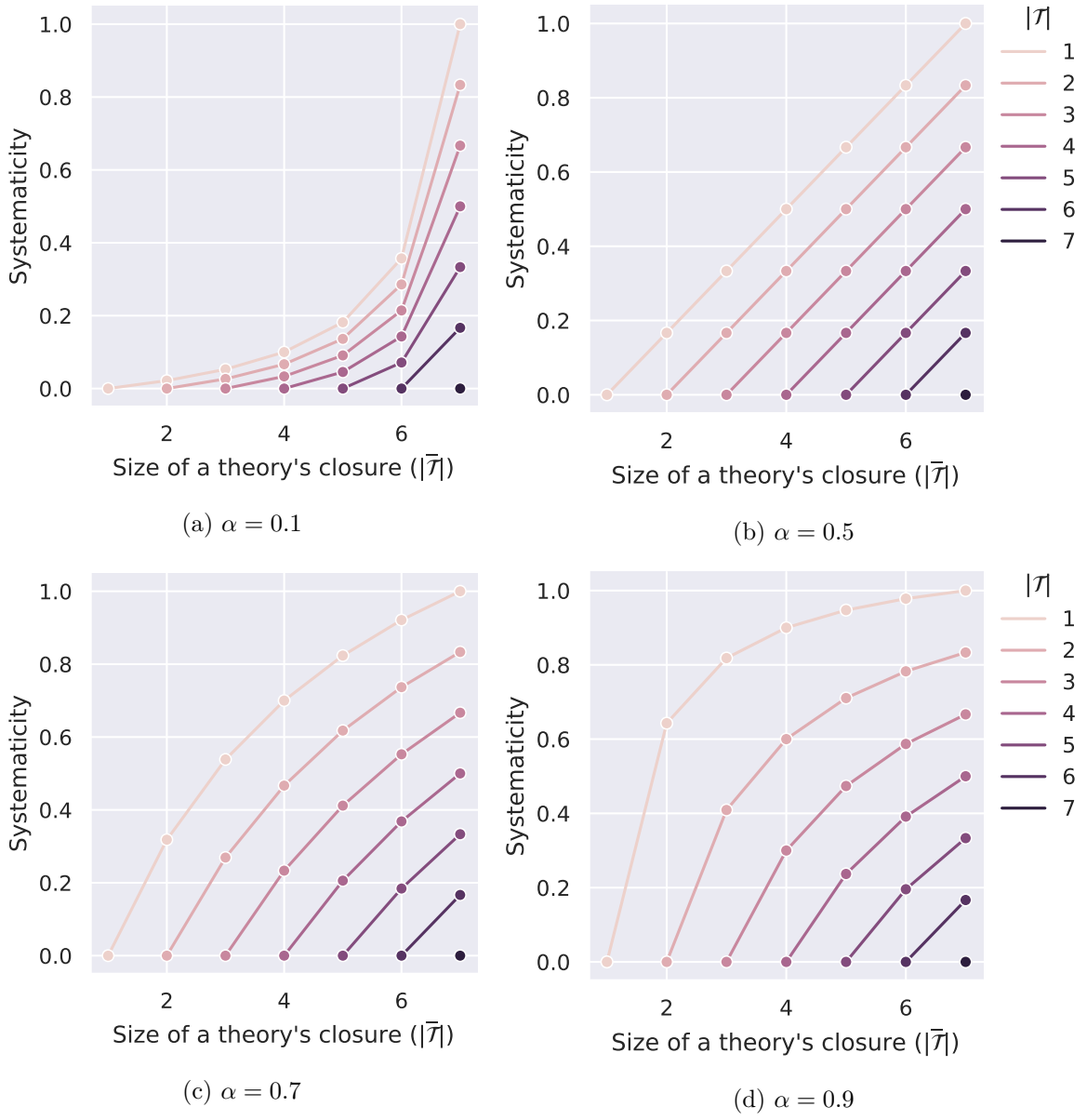
If we want that systematicity decreases with the addition of ad hoc principles (*D4.1*), we must penalize an increase in size more than a decrease in content (i.e., $\alpha > \frac{1}{2}$.) This is illustrated in Figure C.7. For $\alpha = 0.1$ and $\alpha = 0.5$ *D4.1* is violated. If, however, we set $\alpha > 0.5$ (e.g., 0.7 or 0.9), the measure satisfies *D4.1*.

Similarly to S_{ec}, the new measure S_{csw_α} complies with *D3.1 (minimal systematicity)*. They do so in a very specific way: The systematicity values for ineffective theories are not only lower bounds for effective theories, but they also receive the same and lowest systematicity value possible, namely 0.

Surely, for a fixed $|\mathcal{T}|$, systematicity should be minimised for $|\overline{\mathcal{T}}| = |\mathcal{T}|$ and vice versa. However, it is not clear that all cases of $|\overline{\mathcal{T}}| = |\mathcal{T}|$ should have equal systematicity of 0. Especially if we conceive systematicity to be a weighted combination of simplicity and content, we might think that cases of larger $|\overline{\mathcal{T}}| = |\mathcal{T}|$ are better or worse than cases of smaller ones. In particular, if simplicity has more weight than content ($\alpha > 0.5$), then smaller ones should be (a little) more systematic than larger ones (because they are simpler).

This suggests an alternative normalization. For $\alpha > 0.5$, the worst case would be $|\overline{\mathcal{T}}| = |\mathcal{T}| = n$ (minimal simplicity). Plugging this into the penalty function Equation C.2 gives us a normalizing denominator of $\alpha \cdot (n-1)$. For $\alpha < 0.5$, the worst case would be $|\overline{\mathcal{T}}| = |\mathcal{T}| = 1$ (minimal content). Plugging this into the penalty function gives us the normalizing denominator $(1 - \alpha) \cdot (n - 1)$. The denominator that covers both cases is $(|\alpha - 0.5| + 0.5) \cdot (n - 1)$.

To better distinguish the resulting alternative measure from S_{sw_α}, we rename the parameter α to β. This gives:



(a) $\alpha = 0.1$

(b) $\alpha = 0.5$

(c) $\alpha = 0.7$

(d) $\alpha = 0.9$

Figure C.7: Content-simplicity weigthed systematicity (alpha) of theories in dependence of their size and closure's size.

$$S_{sw_\beta}(\mathcal{T}) = 1 - \frac{\beta \cdot (|\mathcal{T}| - 1) + (1 - \beta) \cdot (n - |\overline{\mathcal{T}}|)}{(|\beta - 0.5| + 0.5) \cdot (n - 1)}$$

Similar to S_{sw_α}, S_{sw_β} satisfies *D1*, *D2*. The desiderata *D3.1* and *D4.1* are satisfied for certain valus of α (in our case for $\alpha == 0.525$) as illustrated in Figure C.8.

C.2.4 Relative Effective Content Systematicity

The formulation of another solution starts by framing the problem of S_{ec} in the following way: S_{ec} simply measures the number of implied sentences outside the theory's principles (i.e., $|\overline{\mathcal{T}}| - |\mathcal{T}|$). Consequently, S_{ec} cannot distinguish between theories that are expanded by ad hoc principles, that is, principles that do not expand the theory's content besides the added principles.

However, if a theory is expanded by ad hoc principles, its content measured relative to its size ($\frac{|\overline{\mathcal{T}}|-|\mathcal{T}|}{|\overline{\mathcal{T}}|}$) will decrease.

This suggests to measure $\frac{|\overline{\mathcal{T}}|-|\mathcal{T}|}{|\overline{\mathcal{T}}|}$ instead of simply measuring $|\overline{\mathcal{T}}| - |\mathcal{T}|$, e.g., as follows:

$$S(\mathcal{T}) = \frac{|\overline{\mathcal{T}}| - |\mathcal{T}|}{|\mathcal{T}|(n-1)}$$

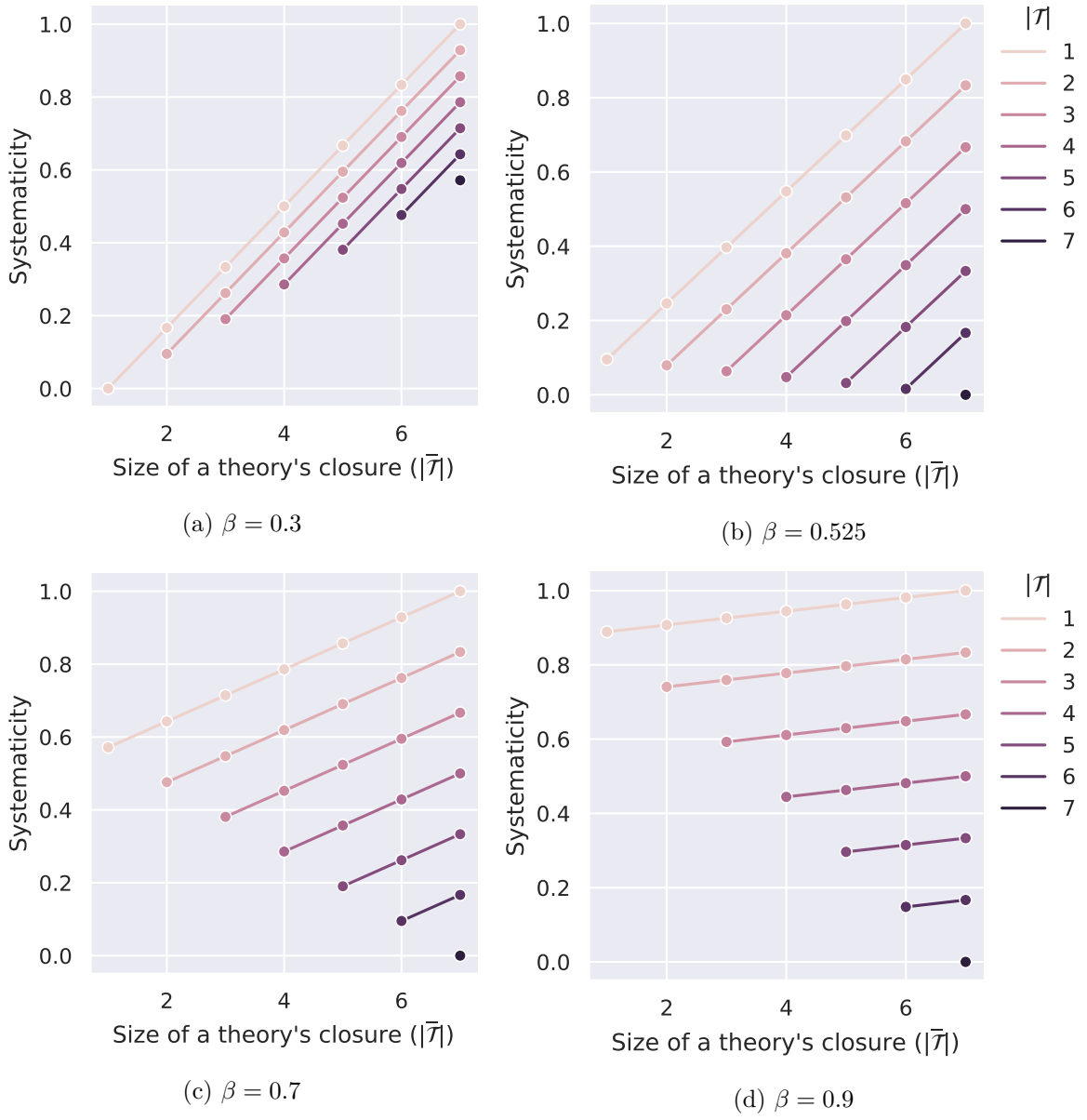
Alternatively, we can conceptualize $\frac{|\overline{\mathcal{T}}|-|\mathcal{T}|}{|\mathcal{T}|}$ as a multiplicative correction factor for S_{ec} which can lead to the following:

$$S(\mathcal{T}) = S_{ec}\frac{|\overline{\mathcal{T}}| - |\mathcal{T}|}{|\mathcal{T}|(n-1)}$$

This quadratic form might, however, decrease $S(\mathcal{T})$ unnecessarily, which motivates us to take the square root of the latter expression:

$$S_{rec}(\mathcal{T}) := \sqrt{S_{ec}\frac{|\overline{\mathcal{T}}| - |\mathcal{T}|}{|\mathcal{T}|(n-1)}} = \frac{|\overline{\mathcal{T}}| - |\mathcal{T}|}{\sqrt{|\mathcal{T}|}(n-1)}$$

Figure C.9 illustrates that $S_{rec}(\mathcal{T})$ satisfies *D1*, *D2*, *D3.1* and *D4.1*.



(a) $\beta = 0.3$

(b) $\beta = 0.525$

(c) $\beta = 0.7$

(d) $\beta = 0.9$

Figure C.8: Content-simplicity weighted systematicity (beta) of theories in dependence of their size and closure's size .

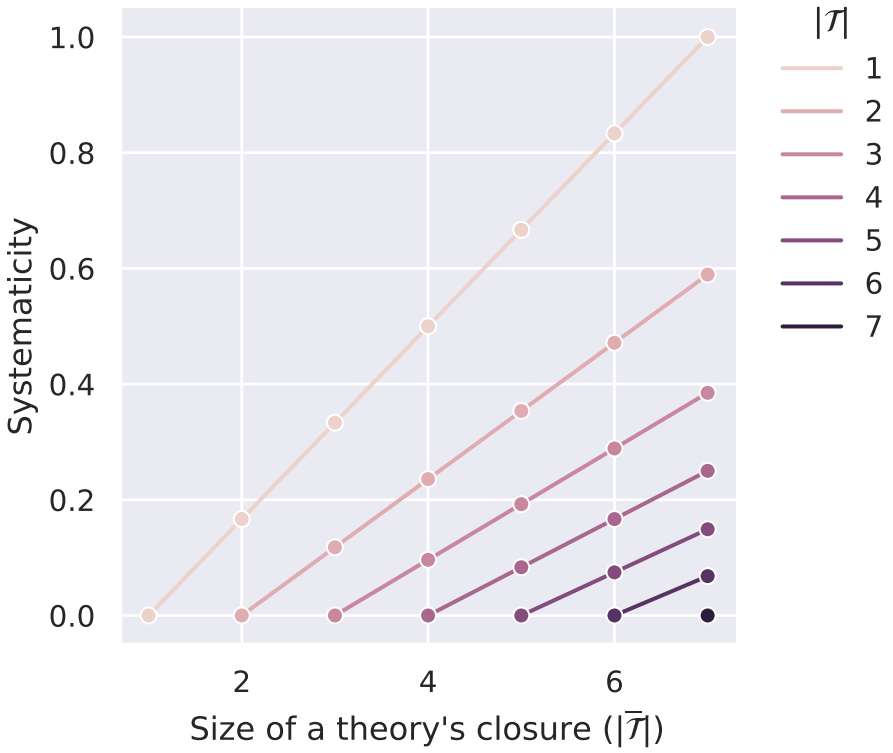


Figure C.9: Relative effective content systematicity of theories in dependence of their size and closure's size.

C.3 Sigma-Based Systematicity Measures

All simple systematicity measures are not able to account for *D5 (internal connectedness)* and *D6 (external connectedness)* for the simple reason that they evaluate systematicity based on $|\mathcal{T}|$ and $|\overline{\mathcal{T}}|$ alone. The problem, in particular, is that the dialectical closure as defined by $\overline{\mathcal{T}} = \{s \in \mathcal{S} | \mathcal{T} \models_\tau s\}$ will miss those dialectical implications that are relevant for *D5* and *D6*. The dialectical closure $\overline{\mathcal{T}}$ enumerates only atomic sentences as implications. While it is clear that arbitrary conjunctions of these atomic sentences are also implications, $\overline{\mathcal{T}}$ is blind to other complex implications of $\mathcal{T}$. It can, in particular, not distinguish between theories $\mathcal{T}_1$ and $\mathcal{T}_2$ for which $\overline{\mathcal{T}_1} \neq \overline{\mathcal{T}_2}$ but which differ with respect to certain disjunctions implied by the theories (i.e., extensional if-then clauses or "conditional implications").

If we want to account for *D5* and *D6*, we have to use a more ambitious concept of *content*. In the following, we will draw on the theory of dialectical structures (Betz 2013) to explicate such a notion of *content*.

The *inferential density* of a dialectical structure τ "can be understood as measure of the inferential constraints encoded in τ" (Betz 2013, 44) and is defined as

$$D(\tau) = \frac{n - lg(\sigma)}{n}$$

with σ being the number of complete and dialectically consistent positions on a dialectical structure τ and $2n$ the size of the sentence pool $\mathcal{S}$.

A *position* is a set of sentences from $\mathcal{S}$ (e.g., the commitments of an epistemic state in our RE model). A dialectical structure will render some of these positions dialectically inconsistent. For instance, an argument with one premise s_1 and the conclusion s_2 renders the position $\{s_1, \neg s_2\}$ dialectically inconsistent.

A *complete position* is a position that includes for each $s \in \mathcal{S}$ either s or $\neg s$. Hence, complete positions do not include flat contradictions (s and $\neg s$), i.e. they are minimally consistent. If a dialectical structure is ineffective, and thus does not render any position dialectically inconsistent, then there are 2^n complete and consistent positions. In this case, $D(\tau) = 0$. On the other hand, if τ allows for exactly one complete and consistent position, and hence renders all other complete positions dialectically inconsistent, then we have $D(\tau) = 1$.

It is straightforward to generalize the concept of inferential density to a notion of content. The inferential density $D(\tau)$ measures how many complete positions are dialectically inconsistent given the dialectical structure alone. We can now ask which complete positions are rendered *additionally* inconsistent if we further assume the truth of sentences from a theory $\mathcal{T}$. In other words, if $\sigma_\mathcal{T}$ is the number of complete consistent positions that extend a theory $\mathcal{T}$, the term $\sigma - \sigma_\mathcal{T}$ can be taken to measure the (σ-based) content size $|C_\sigma(\mathcal{T})|$ of a theory. Proper normalization leads to the following:

$$|C_\sigma(\mathcal{T})| = \frac{lg(\sigma - \sigma_\mathcal{T} + 1)}{n} \tag{C.3}$$

If $\sigma = 2^n$ (minimal inferential density) and $\sigma_\mathcal{T} = 1$ (maximal content), $|C_\sigma(\mathcal{T})| = 1$. If, on the other hand, the theory cannot render anything inconsistent that is not already inconsistent by τ alone (i.e., $\sigma = \sigma_\mathcal{T}$), $|C_\sigma(\mathcal{T})| = 0$.

The more implications $\mathcal{T}$ has, the more complete positions are (additionally) rendered dialectically inconsistent. In this way, the expression Equation C.3 is a natural generalization of $|\overline{\mathcal{T}}|$.[6]

C.3.1 Generalizing Relative Effective Content Systematicity

There are surely different possibilities to introduce systematicity measures based on the (σ-based) content size $|C_\sigma(\mathcal{T})|$. Here, we discuss but one measure, which is based on the considerations we used to devise the measure S_{rec} in Section C.2.4. The basic idea of this measure was to take the effective content size $|\overline{\mathcal{T}} \setminus \mathcal{T}|$ in relation to the size of the theory.

Hence, the first step is to find a generalization of the expression $|\overline{\mathcal{T}} \setminus \mathcal{T}|$. Similar to S_{rec}, we are interested in what a theory inferentially accomplishes besides implying its principles. To that end, we might consider those positions that are complete and dialectically consistent outside the domain of the theory. That is, we consider a restriction of the sentence pool $\mathcal{S}$ to those sentences that are neither principles nor negations of a theory's principles.

More formally, let $\mathcal{S}_\mathcal{T}$ be the domain of $\mathcal{T}$, and let $\mathcal{S} \setminus \mathcal{T} = \mathcal{S} \setminus \mathcal{S}_\mathcal{T}$ be the domain outside the principles of $\mathcal{T}$, $2m$ the size of the restricted sentence pool and $\sigma_\mathcal{T}^{\mathcal{S}\setminus\mathcal{T}}$ the number of positions that are dialectically consistent given $\mathcal{T}$ and complete on the restricted domain $\mathcal{S} \setminus \mathcal{T}$. Then, we can define the *effective content size* on $\mathcal{S} \setminus \mathcal{T}$ as

$$|C_\sigma(\mathcal{T}, \mathcal{S} \setminus \mathcal{T})| = \frac{lg(\sigma^{\mathcal{S}\setminus\mathcal{T}} - \sigma_\mathcal{T}^{\mathcal{S}\setminus\mathcal{T}} + 1)}{m}. \tag{C.4}$$

Similar to S_{rec}, systematicity should measure the effective content size relative to the size of the theory—that is, something like:

$$S \propto \frac{lg(\sigma^{\mathcal{S}\setminus\mathcal{T}} - \sigma_\mathcal{T}^{\mathcal{S}\setminus\mathcal{T}} + 1)}{|\mathcal{T}| \cdot m}$$

What remains is a proper normalization.

[6] In fact, Equation C.3 is more akin to $|\overline{\mathcal{T}}|$ minus the amount of τ truths (i.e., sentences that are tautologically true with respect to the dialectical structure τ).

Maximal effective content is reached by singleton theories that render all but one position inconsistent. Under this assumption we have $\sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}} = 1$ and $|\mathcal{T}| = 1$. The value of $\sigma^{\mathcal{S}\setminus\mathcal{T}}$ will depend on the singleton theory in question. One suggestion is, therefore, to take $max(\{\sigma^{\mathcal{S}\setminus\{s\}})|s \in \mathcal{S}\}$ to normalize the measure.

For simplicity, we will use another estimation. $\sigma^{\mathcal{S}\setminus\mathcal{T}}$ will be maximal if the dialectical structure is silent on the domain $\mathcal{S} \setminus \mathcal{T}$; that is, if it doesn't render anything dialectically inconsistent on $\mathcal{S} \setminus \mathcal{T}$. In this case, $\sigma^{\mathcal{S}\setminus\mathcal{T}}$ will be 2^{n-1} for singleton theories. This motivates the following normalization:[7]

$$S_{grec}(\mathcal{T}) = \frac{lg(\sigma^{\mathcal{S}\setminus\mathcal{T}} - \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}} + 1)}{|\mathcal{T}| \cdot (n-1)} \tag{C.5}$$

How does this measure perform with respect to the different desiderata?

For the simple measures, we equated content with $\overline{\mathcal{T}}$. Since we adopted a more ambitious concept of content for the generalized measure, we have to assess its performance with respect to this explication of content.

The desiderata *D1 (content)* and *D2 (simplicity)* are trivially satisfied. The numerator of Equation C.5 is proportional to the size of the generalized content; the denominator is proportional to the theory size. Accordingly, if we keep the size of the theory constant, systematicity increases with increasing content (*D1*). Similarly, if we keep the theory's content constant, systematicity increases with a decrease in theory size (*D2*). Figure Figure C.10 illustrates this behaviour.[8]

Ineffective theories do not imply anything besides their principles that is not already tautologically true (with respect to τ). Accordingly, they do not render any positions inconsistent on $\sigma^{\mathcal{S}\setminus\mathcal{T}}$ that are not already dialectically inconsistent according to τ alone. Hence, we have $\sigma^{\mathcal{S}\setminus\mathcal{T}} = \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}}$, which implies that $S_{grec}(\mathcal{T}) = 0$ for ineffective theories. Effective theories, on the other hand, do imply something additional on $\sigma^{\mathcal{S}\setminus\mathcal{T}}$. Hence, we have $\sigma^{\mathcal{S}\setminus\mathcal{T}} > \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}}$ and accordingly $S_{grec}(\mathcal{T}) > 0$ for effective theories. Taken together, this implies that systematicity values of ineffective theories are lower bounds for those of effective theories (*D3.1*).

It is difficult to assess the desideratum *D4.1* visually (as we did with the simple measures) since we cannot identify ad hoc extensions of theories in Figure C.10. However, similar to *D3.1*, we can provide a proof that S_{grec} conforms to *D4.1*.

[7] Admittedly, S_{grec} will under this construction never reach the value one, because $\sigma^{\mathcal{S}\setminus\mathcal{T}} = 2^{n-1}$ means that the theory won't imply anything on $\mathcal{S} \setminus \mathcal{T}$. However, S_{grec} will still have maximal systematicity values for singleton theories that have maximal content on $\mathcal{S} \setminus \mathcal{T}$.

[8] The figure is plotted based on a data set of 100 randomly generated dialectical structures and all possible theory candidates for each τ. This dataset is not needed to plot the function S_{grec}. However, it contains all information of τ-theory pairs to assess σ-based systematicity measures in detail (e.g., it contains for each pair $\sigma^{\mathcal{S}\setminus\mathcal{T}}$ and $\sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}}$).

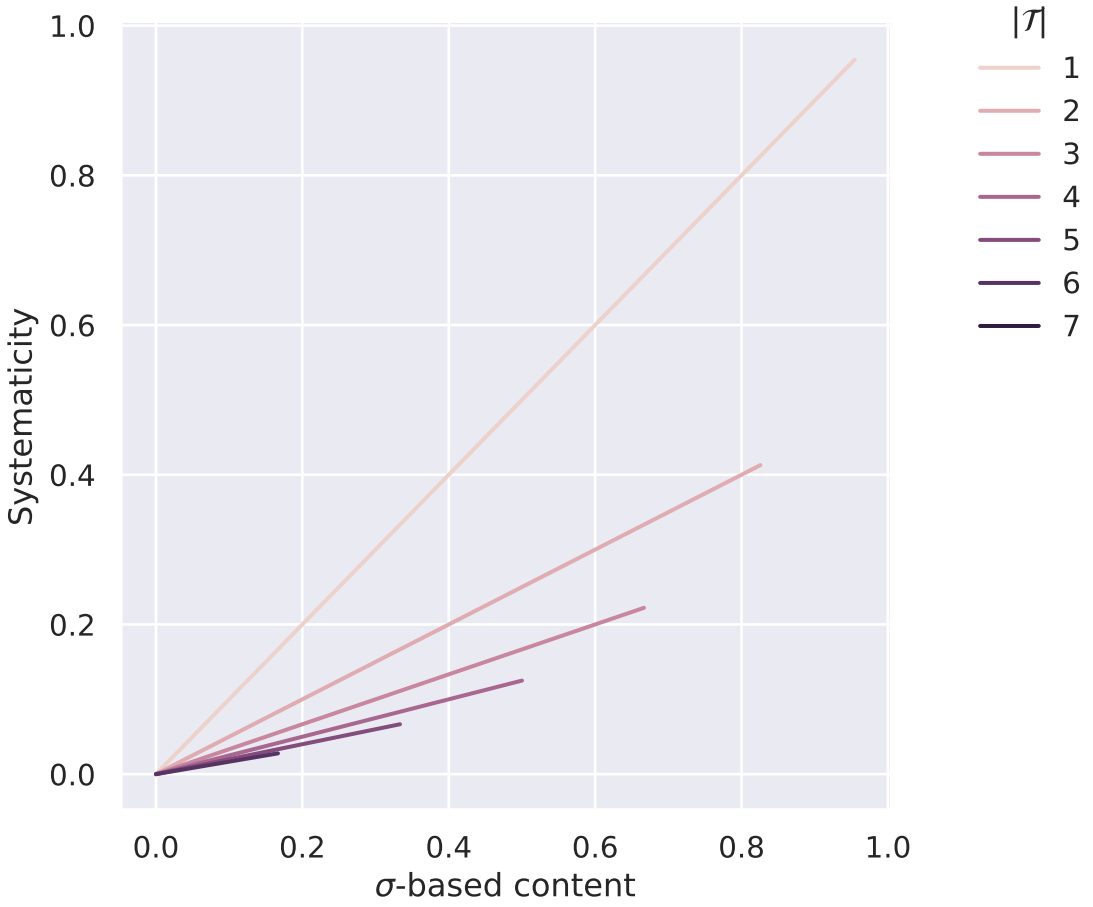


Figure C.10: Generalized relative effective content size systematicity of theories in dependence of their size and closure's size.

Lemma C.1. *The generalized relative effective content systematicity satisfies D4.1.*

Proof. We have to show that $S_{grec}(\mathcal{T}^*) < S_{grec}(\mathcal{T})$ if $\mathcal{T}^*$ is an extension of $\mathcal{T}$ with mere ad hoc principles. So let us assume that $\mathcal{T}^*$ is the result of adding an ad hoc principle $p \in \mathcal{S}$ to a theory $\mathcal{T}$.

We have to show that

$$\frac{lg(\sigma^{\mathcal{S}\setminus\mathcal{T}^*} - \sigma_{\mathcal{T}^*}^{\mathcal{S}\setminus\mathcal{T}^*} + 1)}{|\mathcal{T}^*| \cdot (n-1)} < \frac{lg(\sigma^{\mathcal{S}\setminus\mathcal{T}} - \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}} + 1)}{|\mathcal{T}| \cdot (n-1)}$$

The corresponding comparison for simple systematicity measures is more or less trivial. In these cases, we could simply use that $|\overline{\mathcal{T}} \setminus \mathcal{T}| = |\overline{\mathcal{T}}^* \setminus \mathcal{T}^*|$. Adding one ad hoc principle to a theory increases its closure and size by one. If we compare the change of size and σ-based content, the comparison is not so straightforward any more.

Basically, we have to compare $\sigma^{\mathcal{S}\setminus\mathcal{T}^*} - \sigma_{\mathcal{T}^*}^{\mathcal{S}\setminus\mathcal{T}^*}$ with $\sigma^{\mathcal{S}\setminus\mathcal{T}} - \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}}$. Clearly, $\sigma^{\mathcal{S}\setminus\mathcal{T}^*} \leq \sigma^{\mathcal{S}\setminus\mathcal{T}}$ (since $\mathcal{S} \setminus \mathcal{T}^* \subset \mathcal{S} \setminus \mathcal{T}$). Additionally, we can use the definition of ad hoc principles: Adding ad hoc principles to a theory $\mathcal{T}$ does not do anything in addition to $\mathcal{T}$ on $\mathcal{S} \setminus \mathcal{T}^*$. Hence, $\sigma_{\mathcal{T}^*}^{\mathcal{S}\setminus\mathcal{T}^*} = \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}^*}$. Considering this equation, we have to compare $\sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}^*}$ with $\sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}}$ and, again, we have $\sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}^*} \leq \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}}$. But this simple estimation does not help to get us any further with comparing $\sigma^{\mathcal{S}\setminus\mathcal{T}^*} - \sigma_{\mathcal{T}^*}^{\mathcal{S}\setminus\mathcal{T}^*}$ and $\sigma^{\mathcal{S}\setminus\mathcal{T}} - \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}}$.

What we need is a more precise estimation in terms of $\sigma^{\mathcal{S}\setminus\mathcal{T}^*} + a = \sigma^{\mathcal{S}\setminus\mathcal{T}}$ and $\sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}^*} + b = \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}}$. In other words, we need to know the extent to which $\sigma_{(\mathcal{T})}^{\mathcal{S}\setminus\mathcal{T}}$ increases when further restricting the sentence pool.

Let's start with $\sigma^{\mathcal{S}\setminus\mathcal{T}^*}$ and $\sigma^{\mathcal{S}\setminus\mathcal{T}}$. We will search for an additive expression for both terms. Let Γ be the set of all complete and consistent positions on τ (hence, $\sigma = |\Gamma|$). In analogy to $\sigma^{\mathcal{S}'}$, we will define $\Gamma^{\mathcal{S}'}$ as the set of consistent positions that are complete on the subdomain $\mathcal{S}' \subset \mathcal{S}$. More formally, we can define:

$$\Gamma^{\mathcal{S}'} = \{\mathcal{A} \cap \mathcal{S}' | \mathcal{A} \in \Gamma\}$$

We will now partition $\Gamma^{\mathcal{S}\setminus\mathcal{T}^*}$. Since, $\mathcal{S} \setminus \mathcal{T}^* \subset \mathcal{S} \setminus \mathcal{T}$, we have

$$\Gamma^{\mathcal{S}\setminus\mathcal{T}^*} = \{\mathcal{A} \cap \mathcal{S} \setminus \mathcal{T}^* | \mathcal{A} \in \Gamma^{\mathcal{S}\setminus\mathcal{T}}\} \tag{C.6}$$

In other words, elements in $\Gamma^{\mathcal{S}\setminus\mathcal{T}^*}$ result from reducing elements in $\Gamma^{\mathcal{S}\setminus\mathcal{T}}$ to the domain outside $\mathcal{T}^*$. Since the domains $\mathcal{S} \setminus \mathcal{T}$ and $\mathcal{S} \setminus \mathcal{T}^*$ only differ with respect to the principle p and its negation (i.e., $\mathcal{S} \setminus \mathcal{T} - \mathcal{S} \setminus \mathcal{T}^* = \{p, \neg p\}$), there are three (exclusive) possibilities of how elements from $\Gamma^{\mathcal{S}\setminus\mathcal{T}}$ are mapped to elements from $\Gamma^{\mathcal{S}\setminus\mathcal{T}^*}$: For all $A \in \Gamma^{\mathcal{S}\setminus\mathcal{T}^*}$ either

1. $A \cup \{p\} \in \Gamma^{\mathcal{S}\setminus\mathcal{T}}$ and $A \cup \{\neg p\} \notin \Gamma^{\mathcal{S}\setminus\mathcal{T}}$, or
2. $A \cup \{p\} \notin \Gamma^{\mathcal{S}\setminus\mathcal{T}}$ and $A \cup \{\neg p\} \in \Gamma^{\mathcal{S}\setminus\mathcal{T}}$, or
3. $A \cup \{p\} \in \Gamma^{\mathcal{S}\setminus\mathcal{T}}$ and $A \cup \{\neg p\} \in \Gamma^{\mathcal{S}\setminus\mathcal{T}}$.

The corresponding sets are denoted by $\Gamma_1^{\mathcal{S}\setminus\mathcal{T}^*}$, $\Gamma_2^{\mathcal{S}\setminus\mathcal{T}^*}$ and $\Gamma_3^{\mathcal{S}\setminus\mathcal{T}^*}$ and represent a partitioning of $\Gamma^{\mathcal{S}\setminus\mathcal{T}^*}$:

$$\Gamma^{\mathcal{S}\setminus\mathcal{T}^*} = \Gamma_1^{\mathcal{S}\setminus\mathcal{T}^*} \cup \Gamma_2^{\mathcal{S}\setminus\mathcal{T}^*} \cup \Gamma_3^{\mathcal{S}\setminus\mathcal{T}^*}$$

Similar to the definition of $\sigma_{\mathcal{T}}$, let $\Gamma_{\mathcal{T}}$ the set of complete positions that extend $\mathcal{T}$. Using this definition, we can partition $\Gamma^{\mathcal{S}\setminus\mathcal{T}}$ into

$$\Gamma^{\mathcal{S}\setminus\mathcal{T}} = \Gamma_{\{p\}}^{\mathcal{S}\setminus\mathcal{T}} \cup \Gamma_{\{\neg p\}}^{\mathcal{S}\setminus\mathcal{T}}$$

We can now use the above defined sets to rewrite $\Gamma_{\{p\}}^{\mathcal{S}\setminus\mathcal{T}}$ and $\Gamma_{\{\neg p\}}^{\mathcal{S}\setminus\mathcal{T}}$ in the following way:

$$\Gamma_{\{p\}}^{\mathcal{S}\setminus\mathcal{T}} = \left(\Gamma_1^{\mathcal{S}\setminus\mathcal{T}^*} \cup \{p\}\right) \cup \left(\Gamma_3^{\mathcal{S}\setminus\mathcal{T}^*} \cup \{p\}\right)$$

$$\Gamma_{\{\neg p\}}^{\mathcal{S}\setminus\mathcal{T}} = \left(\Gamma_2^{\mathcal{S}\setminus\mathcal{T}^*} \cup \{\neg p\}\right) \cup \left(\Gamma_3^{\mathcal{S}\setminus\mathcal{T}^*} \cup \{\neg p\}\right)$$

This leads to

$$\sigma^{\mathcal{S}\setminus\mathcal{T}^*} = \sigma_1^{\mathcal{S}\setminus\mathcal{T}^*} + \sigma_2^{\mathcal{S}\setminus\mathcal{T}^*} + \sigma_3^{\mathcal{S}\setminus\mathcal{T}^*}$$

and

$$\sigma^{\mathcal{S}\setminus\mathcal{T}} = \sigma_1^{\mathcal{S}\setminus\mathcal{T}^*} + \sigma_2^{\mathcal{S}\setminus\mathcal{T}^*} + 2\sigma_3^{\mathcal{S}\setminus\mathcal{T}^*}$$

and hence

$$\sigma^{\mathcal{S}\setminus\mathcal{T}} = \sigma^{\mathcal{S}\setminus\mathcal{T}^*} + \sigma_3^{\mathcal{S}\setminus\mathcal{T}^*} \tag{C.7}$$

An analogical partitioning of $\Gamma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}^*}$ and $\Gamma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}}$ yields

$$\sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}} = \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}^*} + (\sigma_{\mathcal{T}})_3^{\mathcal{S}\setminus\mathcal{T}^*} \tag{C.8}$$

We will now use Equation C.7 and Equation C.8 to show that $S_{grec}(\mathcal{T}^*) < S_{grec}(\mathcal{T})$.

Clearly, $(\sigma_{\mathcal{T}})_3^{\mathcal{S}\setminus\mathcal{T}^*} \leq \sigma_3^{\mathcal{S}\setminus\mathcal{T}^*}$. Equation C.7 and Equation C.8 can now be used to deduce

$$\sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}} - \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}^*} \leq \sigma^{\mathcal{S}\setminus\mathcal{T}} - \sigma^{\mathcal{S}\setminus\mathcal{T}^*}$$

Since $\mathcal{T}^*$ is an ad hoc extension of $\mathcal{T}$, we have $\sigma_{\mathcal{T}^*}^{\mathcal{S}\setminus\mathcal{T}^*} = \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}^*}$, which leads to

$$\sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}} - \sigma_{\mathcal{T}^*}^{\mathcal{S}\setminus\mathcal{T}^*} \leq \sigma^{\mathcal{S}\setminus\mathcal{T}} - \sigma^{\mathcal{S}\setminus\mathcal{T}^*}$$

which can be rewritten as

$$\frac{lg(\sigma^{\mathcal{S}\setminus\mathcal{T}^*} - \sigma_{\mathcal{T}^*}^{\mathcal{S}\setminus\mathcal{T}^*} + 1)}{(n-1)} \leq \frac{lg(\sigma^{\mathcal{S}\setminus\mathcal{T}} - \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}} + 1)}{(n-1)} \tag{C.9}$$

Since $\frac{|\mathcal{T}^*|}{|\mathcal{T}|} > 1$ we have also

$$\frac{lg(\sigma^{\mathcal{S}\setminus\mathcal{T}} - \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}} + 1)}{(n-1)} < \frac{|\mathcal{T}^*|}{|\mathcal{T}|}\frac{lg(\sigma^{\mathcal{S}\setminus\mathcal{T}} - \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}} + 1)}{(n-1)} \tag{C.10}$$

Using both estimations Equation C.9 and Equation C.10, we arrive at:

$$\frac{lg(\sigma^{\mathcal{S}\setminus\mathcal{T}^*} - \sigma_{\mathcal{T}^*}^{\mathcal{S}\setminus\mathcal{T}^*} + 1)}{|\mathcal{T}^*|(n-1)} < \frac{lg(\sigma^{\mathcal{S}\setminus\mathcal{T}} - \sigma_{\mathcal{T}}^{\mathcal{S}\setminus\mathcal{T}} + 1)}{|\mathcal{T}|(n-1)}$$

Hence, $S_{grec}(\mathcal{T}^*) < S_{grec}(\mathcal{T})$ if $\mathcal{T}^*$ is an ad hoc extension of $\mathcal{T}$. This concludes the proof of Lemma C.1. □

How does S_{grec} performs with respect to *D5 (internal connectedness)* and *D6 (external conenectedness)*? Since we did not provide any explications of these desiderata, we only calculated S_{grec} for the given illustrations.

In example Example C.1, we considered two theories $\mathcal{T}_1 = \{1, 2\}$ and $\mathcal{T}_2 = \{7, 8\}$ and expect according to *D5* that $S(\mathcal{T}_2) < S(\mathcal{T}_1)$. However, the calculated values ($S(\mathcal{T}_1) = 0.32$ and $S(\mathcal{T}_2) = 0.40$) yield the exact opposite: $S(\mathcal{T}_2) > S(\mathcal{T}_1)$. Surprisingly, these results can be explained by the same reasoning we used to motivate *D5* (compare Figure C.2). Since the sentences of $S(\mathcal{T}_1)$ (1 and 2) only imply other sentences in their combination and the sentences of $S(\mathcal{T}_2)$ (7 and 8) imply other sentences on their own, there are, for instance, more complete consistent positions given 1 than complete consistent positions given 7. In consequence, $\sigma^{\mathcal{S}\setminus\mathcal{T}_1} < \sigma^{\mathcal{S}\setminus\mathcal{T}_2}$ (25 vs. 49). Additionally, $\sigma_{\mathcal{T}_1}^{\mathcal{S}\setminus\mathcal{T}_1} = \sigma_{\mathcal{T}_2}^{\mathcal{S}\setminus\mathcal{T}_2}$ (4) and accordingly $S(\mathcal{T}_2) > S(\mathcal{T}_1)$.

The same happens in Example C.2 (Figure C.3). There we expected $S(\mathcal{T}_2) < S(\mathcal{T}_1)$ for the given theories. However, the calculation of S_{grec} yields: $S(\mathcal{T}_2) > S(\mathcal{T}_1)$ (0.20 vs. 0.14)—again due to $\sigma^{\mathcal{S}\setminus\mathcal{T}_1} < \sigma^{\mathcal{S}\setminus\mathcal{T}_2}$ (9 vs. 16).

In consequence, principles working together is a disadvantage in terms of systematicity measured this way. The examples were intentionally constructed to yield $\sigma_{\mathcal{T}_1}^{\mathcal{S}\setminus\mathcal{T}_1} = \sigma_{\mathcal{T}_2}^{\mathcal{S}\setminus\mathcal{T}_2}$ since we wanted to compare theories that differ to each other only in whether their principles work together. However, the only remaining relevant quantity in $(\sigma^{\mathcal{S}\setminus\mathcal{T}})$ will induce systematicity values in conflict with *D5*.

Example C.3 (Figure C.4) was used to motivate *D6 (external connectedness)*. According to the formulated intuitions, everything else being equal, a theory's systematicity should exceed another's if the former implies more in combination with other sentences than the latter. The simple measures cannot satisfy *D6* since they are insensitive to the non-trivial content (i.e., the content outside $\overline{\mathcal{T}}$). On the other hand, the effective content size $|C_\sigma(\mathcal{T}, \mathcal{S}\setminus\mathcal{T})|$ was conceptualized to account for these implications. Accordingly, it is not surprsing that S_{grec} satisfies *D6*.[9] In the example, we expected that $S(\mathcal{T}_2) < S(\mathcal{T}_1)$, which is confirmed by the corresponding calculations (0 vs. 0.78).

C.4 Conclusion

We motivated the desiderata *D1-D6* by arguing that the systematicity measure S_{BBB} used in Beisbart, Betz, and Brun (2021) has some shortcomings and by alluding to some general intuitions concerning the concept of systematicity (Section C.1). We moved on to motivate some alternative measures and discussed their advantages and disadvantages in terms of *D1-D6* (see Table C.1 for an overview).

Simple systematicity measures calculate their values based on the two quantities $|\mathcal{T}|$ and $|\overline{\mathcal{T}}|$. Accordingly, all simple measures cannot account for *D5* and *D6*, which demands the consideration of additional properties of theories.

S_{BBB} does not (fully) satisfy *D1* and does not satisfy *D3.1*. The most simple adaption of S_{BBB} (S_{mm}) satisfies *D1-D4.1*. The measures S_{ec} and S_{ec^2} are also able to fix the shortcomings of S_{BBB} but do not satisfy *D4.1*. We suggested three adaptions of S_{ec} that satisfy *D1-D4.1*, two of which incorporate an additional parameter to model the balancing between content and simplicity.

Sigma-based measures draw on a more sophisticated notion of content (Section C.3), which can be used to devise additional systematicity measures. We suggested one systematicity measure that is able to account for *D1-D4.1* and *D6* but which does not satisfy *D5* (Section C.3.1).

[9] At least, if *'everything else being equal'* includes $\sigma^{\mathcal{S}\setminus\mathcal{T}}$.

Systematicity measure	D1	D2	D3.1	D4.1	D5	D6
Standard measure (S_{BBB})	☒	☑	☒	☑	☒	☒
Minimal mutation systematicity (S_{mm})	☑	☑	☑	☑	☒	☒
Effective content systematicity (S_{ec})	☑	☑	☑	☒	☒	☒
Quadratic effective content systematicity (S_{ec^2})	☑	☑	☑	☒	☒	☒
Weighted systematicity (S_{csw_α})	☑	☑	☑	☑	☒	☒
Weighted systematicity (S_{csw_β})	☑	☑	☑	☑	☒	☒
Relative effective content systematicity (S_{rec})	☑	☑	☑	☑	☒	☒
Generalized effective content systematicity (S_{grec})	☑	☑	☑	☑	☒	☑

Table C.1: Overview of how the different measures conform to the suggested requirements *content (D1), simplicity (D2), minimal systematicity (D3.1), non-ad-hocness (D4.1), internal connectedness (D5) and external connectedness (D6).*

The described results are preliminary in that they do not prescribe to replace the measure S_{BBB}.

First of all, we did not provide any simulation results of model variants using these alternative measures. Hence, we do not know how these model variants perform with respect to the described evaluation criteria (Section 2.3).

Second, we formulated but a few intuitions in favour of these desiderata without systematically arguing for them. Hence, it is undecided which of them are important (and to what extent).

Finally, even if these desiderata are important, there are other possibilities to account for them. Instead of threading them into one complex systematicity measure, they might be separated into different measures that are used to extend the given achievement function. For the measures S_{csw_α} and S_{csw_β} we already suggested that content (*D1*) and simplicity (*D2*) might be weighted against each other. The considerations concerning how S_{grec} performs with respect to *D5* and *D6* also suggest that *D5* and *D6* might be in conflict with each other. Accordingly, making the corresponding balancing explicit (by parametrization) might be appropriate.