Automatically Tailoring Static Analysis to Custom Usage Scenarios

Muhammad Numair Mansur*, Benjamin Mariano†, Maria Christakatis*, Jorge A. Navas‡ and Valentin Wustholz§

*MPI-SWS, Germany
{numair,maria}@mpi-sws.org
†The University of Texas at Austin, USA
bmariano@cs.utexas.edu
‡SRI International, USA
jorge.navas@sri.com
§ConsenSys, Germany
valentin.wustholz@consensys.net

Abstract—In recent years, there has been significant progress in the development and industrial adoption of static analyzers. Such analyzers typically provide a large, if not huge, number of configurable options controlling the precision and performance of the analysis. A major hurdle in integrating static analyzers in the software-development life cycle is tuning their options to custom usage scenarios, such as a particular code base or certain resource constraints.

In this paper, we propose a technique that automatically tailors a static analyzer, specifically an abstract interpreter, to the code under analysis and any given resource constraints. We implement this technique in a framework called TAILOR, which we use to perform an extensive evaluation on real-world benchmarks. Our experiments show that the configurations generated by TAILOR are vastly better than the default analysis options, vary significantly depending on the code under analysis, and most remain tailored to several subsequent code versions.

I. INTRODUCTION

Static analysis inspects code, without running it, in order to prove properties or detect bugs. Typically, static analysis approximates the behavior of the code, for instance, because checking the correctness of most properties is undecidable. Performance is another important reason for this approximation. In general, the closer the approximation is to the actual behavior of the code, the less efficient and the more precise the analysis is, that is, the fewer false positives it reports. For less tight approximations, the analysis often becomes more efficient but less precise.

Recent years have seen tremendous progress in both the development and industrial adoption of static analyzers. Notable successes include Facebook’s Infer [1], [2] and AbsInt’s Astrée [3]. Many popular analyzers, such as these, are based on abstract interpretation [4], a technique that abstracts the concrete program semantics and reasons about its abstraction. In particular, program states are abstracted as elements of abstract domains. Most abstract interpreters offer a wide range of abstract domains that impact the precision and performance of the analysis. For instance, the Intervals domain [5] is typically faster but less precise than Polyhedra [6], which captures linear inequalities among any number of variables.

In addition to the domains, abstract interpreters usually provide a large number of other options, for instance, whether backward analysis should be enabled or how quickly a fixpoint should be reached. In fact, the sheer number of option combinations (over 6M in our experiments) is bound to overwhelm users, especially non-expert ones. To make matters worse, the best option combinations may vary significantly depending on the code under analysis and the resources, such as time or memory, that users are willing to spend.

In light of this, we suspect that most users resort to using the default options that the analysis designer pre-selected for them. However, these options are definitely not suitable for all code. Moreover, they do not adjust to different stages of software development, e.g., running the analysis in the editor should be much faster than running it in a continuous integration (CI) pipeline, which in turn should be much faster than running it prior to a major release. The alternative of enabling the (in theory) most precise analysis can be even worse, since in practice it often runs out of time or memory as we show in our experiments. As a result, the widespread adoption of abstract interpreters is severely hindered, which is unfortunate since they constitute an important class of practical static analyzers.

Our approach. To address this issue, we present the first technique that automatically tailors a generic abstract interpreter to a custom usage scenario. With the term custom usage scenario, we refer to a particular piece of code and specific resource constraints. The key idea behind our technique is to phrase the problem of customizing the abstract-interpretation configuration to a given usage scenario as an optimization problem. Specifically, different configurations are compared using a cost function that penalizes those that prove fewer properties or require more resources. This cost function can guide the configuration search of a wide range of existing optimization algorithms.

We implement our technique in a framework called TAILOR, which configures a given abstract interpreter for a given usage scenario using a given optimization algorithm. As a result, TAILOR enables the abstract interpreter to prove as many properties as possible within the resource limit without
requiring any domain expertise on behalf of the user.

Using TAILOR, we find that tailored configurations vastly outperform the default options pre-selected by the analysis designers. In fact, we show that this is possible even with very simple optimization algorithms. Our experiments also demonstrate that tailored configurations vary significantly depending on the usage scenario—in other words, there cannot be a single configuration that fits all scenarios. Finally, most of the generated configurations remain tailored to several subsequent code versions, suggesting that re-tuning is only necessary after major code changes.

**Contributions.** We make the following contributions:

1) We present the first technique for automatically tailoring abstract interpreters to custom usage scenarios.

2) We implement our technique in a framework called TAILOR.

3) Using a state-of-the-art abstract interpreter with millions of configurations, we show the effectiveness of TAILOR on real-world benchmarks.

**Outline.** In the next section, we give a high-level overview of our technique and framework. Sect. III provides background on the generic architecture of abstract interpreters. Sect. IV describes our technique in detail, and Sect. V presents our experimental evaluation. We discuss related work in Sect. VI and conclude in Sect. VII.

II. OVERVIEW

We now illustrate the workflow and tool architecture of TAILOR and provide examples of its effectiveness.

**Terminology.** In the following, we refer to an abstract domain with all its options (e.g., enabling backward analysis or more precise treatment of arrays etc.) as an *ingredient*.

As discussed earlier, abstract interpreters typically provide a large number of such ingredients. To make matters worse, it is also possible to combine different ingredients into a sequence (which we call a *recipe*) such that more properties are verified than with individual ingredients. For example, a user could configure the abstract interpreter to first use Intervals to verify their editor for instant feedback during code development. This results in the tailored configuration using a completely different abstract interpreter [7] in their editor for instant feedback during code development. This means that the allowed time limit for the analysis is very short, say, 1 sec. Now assume that the code under analysis is a program file\(^1\) of the multimedia processing tool FFmpeg, which is used to evaluate the effectiveness of TAILOR in our experiments. In this file, CRAB checks 45 assertions for common bugs, namely, division by zero, integer overflow, buffer overflow, and use after free.

Analysis of this file with the default CRAB configuration takes 0.35 sec to complete. In this time, CRAB proves 17 assertions and emits 28 warnings about the properties that remain unverified. For this usage scenario, TAILOR is able to tune the abstract-interpreter configuration such that the analysis time is 0.57 sec and the number of verified properties increases by 29% (i.e., 22 assertions are proved). Note that the tailored configuration uses a completely different abstract domain than the one used in the default configuration. As a result, the verification results are significantly better, but the analysis takes slightly longer to complete (although remaining

\(^1\)https://github.com/FFmpeg/FFmpeg/blob/master/libavformat/idcin.c
within the specified time limit). In contrast, enabling the most precise analysis in CRAB verifies 26 assertions but takes over 6 min to complete, which by far exceeds the time limit imposed by the specified usage scenario.

While it takes TAILOR 4.5 sec to find the above configuration, this is time well invested; the configuration can be re-used for several subsequent code versions. In fact, in our experiments, we show that generated configurations can remain tailored for at least up to 50 subsequent commits to a file under version control. Given that changes in the editor are typically much more incremental, we expect that no re-tuning would be necessary at all during an editor session. Re-tuning may be beneficial after major changes to the code under analysis and can happen offline, e.g., between editor sessions, or in the worst case overnight.

As another example, consider the usage scenario where CRAB is integrated in a CI pipeline. In this scenario, users should be able to spare more time for analysis, say, 5 min. Here, let us assume that the analyzed code is a program file \(^2\) of the curl tool for transferring data by URL, which is also used in our evaluation. The default CRAB configuration takes 0.23 sec to run and only verifies 2 out of 33 checked assertions. TAILOR is able to find a configuration that takes 7.6 sec and proves 8 assertions. In contrast, the most precise configuration does not terminate even after 15 min.

Both usage scenarios demonstrate that, even when users have more time to spare, the default configuration cannot take advantage of it to improve the verification results. At the same time, the most precise configuration is completely impractical since it does not respect the resource constraints imposed by these scenarios.

### III. Background: A Generic Abstract Interpreter

Many successful abstract interpreters (e.g., Astrée [3], C Global Surveyor [8], Clousot [9], CRAB [7], IKOS [10], Sparrow [11], and Infer [1]) follow the generic architecture in Fig. 2. In this section, we describe the main components of such a generic abstract interpreter.

**Memory domain.** Analysis of low-level languages such as C and LLVM-bitcode requires reasoning about pointers. It is, therefore, common to design a memory domain [12] that can simultaneously reason about pointer aliasing, memory contents, and numerical relations between them.

**Pointer domains** resolve aliasing between pointers, and **array domains** reason about memory contents. In particular, array domains can reason about individual memory locations (cells), infer universal properties over multiple cells, or both. Typically, reasoning about individual cells trades performance for precision unless there are very few array elements (e.g., [13, 12]). In contrast, reasoning about multiple memory locations (summarized cells) trades precision for performance. In our evaluation, we use **Array smashing** domains [3] that abstract different array elements into a single summarized cell.

**Logico-numerical domains** infer relationships between program and synthetic variables, introduced by the pointer and array domains, e.g., summarized cells. Next, we introduce domains typically used for proving the absence of runtime errors in low-level languages.

**Boolean domains** (e.g., flat Boolean, BDDApron [14]) reason about Boolean variables and expressions. **Non-relational domains** (e.g., Intervals [5], Congruence [15]) do not track relations among different variables, in contrast to **relational domains** (e.g., Equality [16], Zones [17], Octagons [18], Polyhedra [6]). Due to their increased precision, relational domains are typically less efficient than non-relational ones. **Symbolic domains** (e.g., Congruence closure [19], Symbolic constant [20], Term [21]) abstract complex expressions (e.g., non-linear) and external library calls by uninterpreted functions. **Non-convex domains** express disjunctive invariants. For instance, the DisInt domain [9] extends Intervals to a finite disjunction; it retains the scalability of the Intervals domain by keeping only non-overlapping intervals. On the other hand, the Boxes domain [22] captures arbitrary Boolean combinations of intervals, which can often be expensive.

**Fixpoint computation.** To ensure termination of the fixpoint computation, Cousot and Cousot introduce widening [4], [23], which usually incurs a loss of precision. There are three common strategies to reduce this precision loss, which however sacrifice efficiency. First, delayed widening [3] performs a number of initial fixpoint-computation iterations in the hope of reaching a fixpoint before resorting to widening. Second, widening with thresholds [24], [25] limits the number of program expressions (thresholds) that are used when widening. The third strategy consists in applying narrowing [4], [23] a certain number of times.

**Forward and backward analysis.** Classically, abstract interpreters analyze code by propagating abstract states in a **forward** manner. However, abstract interpreters can also perform backward analysis to compute the execution states that lead to an assertion violation. Cousot and Cousot [26], [27] define a **forward-backward refinement** algorithm in which a forward analysis is followed by a backward analysis until no more refinement is possible. The backward analysis uses invariants computed by the forward analysis, while the forward analysis does not explore states that cannot reach an assertion violation based on the backward analysis. This refinement is more precise than forward analysis alone, but it may also become very expensive.

**Intra- and inter-procedural analysis.** An intra-procedural analysis analyzes a function ignoring the information (i.e., call stack) that flows into that function, while an inter-procedural analysis considers all the flows among functions. The former is much more efficient and easy to parallelize, but the latter is usually more precise.

### IV. Our Technique

In this section, we describe the main components of TAILOR in detail; Sects. IV-A, IV-B, IV-C explain the optimization engine, recipe evaluator, and recipe generator from Fig. 1.

---

\(^2\)https://github.com/curl/curl/blob/master/lib/cookie.c
A. Recipe Optimization

Alg. 1 implements the optimization engine. In addition to the code $P$ and the resource limit $r_{\text{max}}$, it also takes as input the maximum length of the generated recipes $l_{\text{max}}$ (i.e., the maximum number of ingredients), a function to generate new recipes $\text{GENERATEREC}$ (i.e., the recipe generator from Fig. 1), and four other parameters, which we explain later.

A tailored recipe is found in two phases. The first phase aims to find the best abstract domain for each ingredient, while the second tunes the remaining analysis settings for each ingredient (e.g., whether backward analysis should be enabled). Parameters $i_{\text{dom}}$ and $i_{\text{set}}$ control the number of iterations of each phase. Note that we start with a search for the best domains since they have the largest impact on the precision and performance of the analysis.

During the first phase, the algorithm initializes the best recipe $\text{rec}_{\text{best}}$ with an initial recipe $\text{rec}_{\text{init}}$ (line 3). The cost of this recipe is evaluated with function $\text{EVAL}$, which implements the recipe-evaluator component from Fig. 1. The subsequent nested loop (line 5) samples a number of recipes, starting with the shortest recipes ($l := 1$) and ending with the longest recipes ($l := l_{\text{max}}$). The inner loop generates $i_{\text{dom}}$ ingredients for each ingredient in the recipe (i.e., $i_{\text{dom}} \cdot l$ total iterations) by invoking function $\text{GENERATEREC}$, and in case a recipe with lower cost is found, it updates the best recipe (lines 9–10). Several optimization algorithms, such as hill climbing and simulated annealing, search for an optimal result by mutating some of the intermediate results. Variable $\text{rec}_{\text{curr}}$ stores intermediate recipes to be mutated, and function $\text{ACCEPT}$ decides when to update it (lines 11–12).

As explained earlier, the purpose of the first phase is to identify the best sequence of abstract domains. The second phase (lines 13–18) focuses on tuning the other settings of the best recipe so far. This is done by randomly mutating the best recipe via $\text{MUTATESettings}$ (line 15), and updating the best recipe if better settings are found (lines 17–18). After exploring $i_{\text{set}}$ random settings, the best recipe is returned to the user (line 19).

B. Recipe Evaluation

The recipe evaluator from Fig. 1 uses a cost function to determine the quality of a fresh recipe with respect to the precision and performance of the abstract interpreter. This design is motivated by the fact that analysis imprecision and inefficiency are among the top pain points for users [28]. Therefore, the cost function depends on the number of generated warnings $w$ (that is, the number of unverified assertions), the total number of assertions in the code $w_{\text{total}}$, the resource consumption $r$ of the analyzer, and the resource limit $r_{\text{max}}$ imposed on the analyzer:

$$\text{cost}(w, w_{\text{total}}, r, r_{\text{max}}) = \begin{cases} w + \frac{r}{r_{\text{max}}} & \text{if } r \leq r_{\text{max}} \\ w_{\text{total}} & \text{otherwise} \end{cases}$$

Note that $w$ and $r$ are measured by invoking the abstract interpreter with the recipe under evaluation. The cost function evaluates to a lower cost for recipes that improve the precision of the abstract interpreter (due to the term $w/w_{\text{total}}$). In case of ties, the term $r/r_{\text{max}}$ causes the function to evaluate to a lower cost for recipes that result in a more efficient analysis. In other words, for two recipes resulting in equal precision, the one with the smaller resource consumption is assigned a lower cost. When a recipe causes the analyzer to exceed the resource limit, it is assigned infinite cost.

C. Recipe Generation

In the literature, there is a broad range of optimization algorithms for different application domains. To demonstrate the generality and effectiveness of TAILOR, we instantiate it with four adaptations of three well-known optimization algorithms, namely random sampling [29], hill climbing (with
regular restarts) [30], and simulated annealing [31], [32]. Here, we describe these algorithms in detail, and in Sect. V, we evaluate their effectiveness.

Before diving into the details, let us discuss the suitability of different kinds of optimization algorithms for our domain. There are algorithms that leverage mathematical properties of the function to be optimized, e.g., by computing derivatives as in Newton’s iterative method. Our cost function, however, is evaluated by running an abstract interpreter, and thus, it is not differentiable or continuous. This constraint makes such analytical algorithms unsuitable. Moreover, evaluating our cost function is expensive, especially for precise abstract domains such as Polyhedra. This makes algorithms that require a large number of samples, such as genetic algorithms, less practical.

Now recall that Alg. 1 is parametric in how new recipes are generated (GENERATEREC) and accepted for further mutations (ACCEPT). Instantiations of these functions essentially constitute our search strategy for a tailored recipe. In the following, we discuss four such instantiations. Note that, in theory, the order of recipe ingredients matters. This is because any properties verified by one ingredient are converted into assumptions for the next, and different assumptions may lead to different verification results. Therefore, all our instantiations are able to explore different ingredient orderings.

**Random sampling.** Random sampling (RS) just generates random recipes of a certain length. Function ACCEPT always returns false as each recipe is generated from scratch, and not as a result of any mutations.

**Domain-aware random sampling.** RS might generate recipes containing two or more abstract domains of comparable precision. For instance, the Octagons domain is typically strictly more precise than Intervals. As a result, a recipe consisting of these domains is essentially equivalent to a recipe containing only Octagons.

Now, assume that we have a partially ordered set (poset) of domains that defines their ordering in terms of precision. An example of such a poset for a particular abstract interpreter is shown in Fig. 3. An optimization algorithm can then leverage this information to reduce the search space of possible recipes. Given such a poset, we therefore define domain-aware random sampling (DARS), which randomly samples recipes that do not contain abstract domains of comparable precision. Again, ACCEPT always returns false.

**Simulated annealing.** Simulated annealing (SA) searches for the best recipe by mutating the current recipe rec new, in Alg. 1. The resulting recipe (rec new), if accepted on line 12, becomes the new recipe to be mutated. Alg. 2 shows an instantiation of GENERATEREC, which mutates a given recipe such that the poset precision constraints are satisfied (i.e., there are no domains of comparable precision). A recipe is mutated either by adding new ingredients with 20% probability or by modifying existing ones with 80% probability (line 2). The probability of adding ingredients is lower to keep recipes short.

When adding a new ingredient (lines 4–5), Alg. 2 calls RANDPOSETLEASTINC, which considers all domains that are incomparable with the domains in the recipe. Given this set, it randomly selects from the domains with the least precision to avoid adding overly expensive domains. When modifying a random ingredient in the recipe (lines 7–16), the algorithm can replace its domain with one of three possibilities: a domain that is immediately more precise (i.e., not transitive in the poset (via POSETGT)), a domain that is immediately less precise (via POSETLT), or an incomparable domain with the least precision (via RANDPOSETLEASTINC). In case the resulting recipe does not satisfy the poset precision constraints, our algorithm retries to mutate the original recipe (lines 17–18).

For simulated annealing, function ACCEPT returns true if the new cost (for the mutated recipe) is less than the current cost. It also accepts recipes whose cost is higher with a certain probability, which is inversely proportional to the cost increase as well as the number of recipes explored so far. In other words, recipes with a small cost increase are likely to be accepted, especially toward the beginning of the exploration.

**Hill climbing.** Our instantiation of hill climbing (HC) performs regular restarts. In particular, it starts with a randomly generated recipe that satisfies the poset precision constraints, generates 10 new valid recipes, and restarts with a random recipe. ACCEPT returns true only if the new cost is lower than the best cost, which is equivalent to the current cost.

V. EXPERIMENTAL EVALUATION

To evaluate our technique, we aim to answer the following research questions:

RQ1: Is our technique effective in finding tailored recipes for different usage scenarios?
RQ2: Are the tailored recipes optimal?
RQ3: How diverse are the tailored recipes?
RQ4: How resilient are the tailored recipes to code changes?

**A. Implementation**

We implemented TAILOR by extending CRAB [7], a parametric framework for modular construction of abstract inter-
interpretation in fact, most static analyses deployed in real usage analysis, our evaluation uses the default intra-procedural analysis. Even though CRA B is a reduced product of Intervals and Congruence, and optionally, reason about array contents using array smashing.

For analyzing these projects using abstract interpretation, our evaluation uses the default inter-procedural analysis; in fact, most static analyses deployed in real usage scenarios are intra-procedural due to time constraints [28].

**B. Benchmark Selection**

For our evaluation, we systematically selected popular and active C projects on GitHub. In particular, we chose the six most starred C repositories with over 300 commits that we could successfully build with the Clang-5.0 compiler. We give a short description of each project in Tab. 2.

For analyzing these projects using abstract interpretation, we needed to introduce properties to be verified. For our purposes, we instrumented these projects with four types of assertions that check for common bugs; namely, division by zero, integer overflow, buffer overflow, and use after free. Introducing assertions to check for runtime errors such as these is common practice in program analysis and verification.

As projects consist of different numbers of files, to avoid skewing the results in favor of a particular project, we randomly and uniformly sampled 20 LLVM-bitcode files from each project, for a total of 120. To ensure that each file was neither too trivial nor too difficult for the abstract interpreter, we used the number of assertions as a complexity indicator and only sampled files with at least 20 assertions and at most 100. Additionally, to guarantee all four assertion types (listed above) were included and avoid skewing the results in favor of a particular assertion type, we required that the sum of assertions for each type was at least 70 across all files—this exact number was largely determined by the benchmarks.

Overall, our benchmark suite of 120 files totals 1346 functions, 5557 assertions (on average 4 assertions per function), and 667927 LLVM instructions (see Tab. 3).

**C. Results**

We now present our experimental results for each research question. We performed all experiments on a 32-core Intel \( \text{\textregistered} \) Xeon \( \text{\textregistered} \) E5-2667 v2 CPU @ 3.30GHz machine with 264GB of memory, running Ubuntu 16.04.1 LTS.

**RQ1: Is our technique effective in finding tailored recipes for different usage scenarios?** We instantiated TAILOR with the four optimization algorithms described in Sect. IV-C: RS, DARS, SA, and HC. We constrained the analysis time to simulate two usage scenarios: 1 sec for instant feedback in the editor, and 5 min for feedback in a CI pipeline. We compare TAILOR with the default recipe (DEF), i.e., the default settings in CRA B as defined by its designer after careful tuning on a large set of benchmarks over the years. DEF uses a combination of two domains, namely, the reduced product of Boolean and Zones. The other default settings are in Tab. 1.

| Project | Functions | Assertions | LLVM Instructions |
|---------|-----------|------------|--------------------|
| CURL    | 306       | 787        | 50541              |
| DARKNET | 130       | 958        | 55847              |
| FFmpeg  | 103       | 888        | 27653              |
| Git     | 218       | 768        | 12230              |
| PHP-SRC | 268       | 1031       | 305943             |
| Redis   | 321       | 1125       | 125639             |
| **Total** | **1346** | **5557**  | **667927**         |

**Table 3: Benchmark characteristics (20 files per project).** The last three columns show the number of functions, assertions, and LLVM instructions in the analyzed files.

---

Table 1: CRA B settings and their possible values as used in our experiments. Default settings are shown in bold.

| Setting                  | Possible Values |
|--------------------------|------------------|
| NUM_DELAY_WIDEN          | {1, 2, 4, 8, 16} |
| NUM_NARROW_ITERATIONS    | {1, 2, 3, 4}     |
| NUM_WIDEN_THRESHOLDS     | {0, 10, 20, 30, 40} |
| BACKWARD_ANALYSIS        | OFF, ON         |
| ARRAY_SMASHING           | OFF, ON         |
| ABSTRACTDOMAINS          | all domains in Fig. 3 |

**Figure 3: Comparing logico-numerical domains in CRA B.** A domain \( d_1 \) is less precise than \( d_2 \) if there is a path from \( d_1 \) to \( d_2 \) going upward, otherwise \( d_1 \) and \( d_2 \) are incomparable.

---

\(^3\text{CRA B is available at https://github.com/seahorn/crab.}\)
For this experiment, we ran TAILOR with each optimization algorithm on the 120 benchmark files, enabling optimization at the granularity of files. Each algorithm was seeded with the same random seed. In Alg. 1, we restrict recipes to contain at most 3 domains ($l_{\text{max}} = 3$) and set the number of iterations for each phase to be 5 and 10 ($i_{\text{dom}} = 5$ and $i_{\text{set}} = 10$).

The results are presented in Fig. 4, which shows the number of assertions that are verified with the best recipe found by each algorithm as well as by the default recipe. All algorithms outperform the default recipe for both usage scenarios, verifying almost twice as many assertions on average.

Fig. 5 gives a more detailed comparison with the default recipe for the time limit of 5 min. In particular, each horizontal bar shows the total number of assertions verified by each algorithm as well as by the default recipe. All algorithms outperform the default recipe for both usage scenarios, verifying almost twice as many assertions on average.

In Fig. 6, we show the total time that each algorithm takes for all iterations. DARS takes longer than all others. This is due to generating more precise recipes thanks to its domain knowledge. Such recipes typically take longer to run but verify more assertions (as in Fig. 4). On average, for all algorithms, TAILOR requires only 30 sec to complete all iterations for the 1-sec timeout and 16 min for the 5-min timeout. As discussed in Sect. II, this tuning time can be spent offline.

**RQ1 takeaway:** TAILOR verifies nearly twice the assertions of the default recipe, regardless of optimization algorithm, timeout, or number of iterations. In fact, even very simple algorithms (such as RS) significantly outperform the default recipe.

**RQ2: Are the tailored recipes optimal?** To check the optimality of the tailored recipes, we compared them with the most precise (and least efficient) CRAB configuration. It uses the most precise domains from Fig. 3 (i.e., $\text{bool, polyhedra, term(int), ric, boxes, and term(disInt)}$) in a recipe of 6 ingredients and assigns the most precise values to all other settings from Tab. 1. We gave a 30-min timeout to this recipe.

For 21 out of 120 files, the most precise recipe ran out of...
memory (264GB). For 86 files, it terminated within 5 min, and for 13, it took longer (within 30 min)—in many cases, this was even longer than TAILOR’s tuning time in Fig. 6. We compared the number of assertions verified by our tailored recipes (which do not exceed 5 min) and by the most precise recipe. For the 86 files that terminated within 5 min, our recipes prove 618 assertions, whereas the most precise recipe proves 534. For the other 13 files, our recipes prove 119 assertions, whereas the most precise recipe proves 98.

Consequently, our (in theory) less precise and more efficient recipes prove more assertions in files where the most precise recipe terminates. Possible explanations for this non-intuitive result are: (1) Polyhedra coefficients may overflow, in which case the constraints are typically ignored by abstract interpreters, and (2) more precise domains with different widening operations may result in less precise results [34], [35].

We also evaluated the optimality of tailored recipes by mutating individual parts of the recipe and comparing to the original. In particular, for each setting in Tab. 1, we tried all possible values and replaced each domain with all other comparable domains in the poset of Fig. 3. For example, for a recipe including zones, we tried octagons, polyhedra, and intervals. In addition, we tried all possible orderings of the recipe ingredients, which in theory could produce different results. We observed whether these changes resulted in a difference in the precision and performance of the analyzer.

Fig. 9 shows the results of this experiment, broken down by setting. Equal (in orange) indicates that the mutated recipe proves the same number of assertions within ±5 seconds of the original. Positive (in green) indicates that it either proves more assertions or the same number of assertions at 5 seconds faster. Negative (in red) indicates that the mutated recipe either proves fewer assertions or the same number of assertions at 5 seconds slower.

The results show that, for our benchmarks, mutating the recipe found by TAILOR rarely led to an improvement. In particular, at least 93% of all mutated recipes were either equal to or worse than the original recipe. In the majority of these cases, mutated recipes are equally good. This indicates that there are many optimal or close-to-optimal solutions and that TAILOR is able to find one of them.

RQ2 takeaway: As compared to the most precise recipe, TAILOR verified more assertions across benchmarks where the most precise recipe terminated. Furthermore, mutating recipes found by TAILOR resulted in improvement only for less than 7% of recipes.

RQ3: How diverse are the tailored recipes? To motivate the need for optimization, we must show that tailored recipes are sufficiently diverse such that they could not be replaced by a well-crafted default recipe. To better understand the characteristics of tailored recipes, we manually inspected all recipes generated by TAILOR.

TAILOR generated recipes of length greater than 1 for 61 files. Out of these, 37 are of length 2 and 24 of length 3. For 77% of generated recipes, NUM_DELAY_WIDEN is not set to the default value of 1. Additionally, 55% of the ingredients enable array smashing, and 32% enable backward analysis.

Fig. 10 shows how often (in percentage) each abstract domain occurs in a best recipe found by TAILOR. We observe that all domains occur almost equally often, with 6 of the 10 domains occurring in between 9% and 13% of recipes. The most common domain was bool at 18%, and the least common was intervals at 4%. We observed a similar distribution of domains even when instrumenting the benchmarks with only one assertion type, e.g., those that check for integer overflow (see Fig. 11).

We also inspected which domain combinations are frequently used in the tailored recipes. One common pattern...
is combinations between bool and numerical domains (18 occurrences). Similarly, we observed 2 occurrences of term(disInt) together with zones. Interestingly, the less powerful variants of combining disInt with zones (3 occurrences) and term(int) with zones (6 occurrences) seem to be sufficient in many cases. Finally, we observed 8 occurrences of polyhedra or octagons with boxes, which are the most precise convex and non-convex domains. Our approach is, thus, not only useful for users, but also for designers of abstract interpreters by potentially inspiring new domain combinations.

**RQ3 takeaway:** The diversity of tailored recipes prevents replacing them with a single default recipe. Over half of the tailored recipes contain more than one ingredient, and ingredients use a variety of domains and their settings.

**RQ4: How resilient are the tailored recipes to code changes?** We expect tailored recipes to be resilient to code changes, i.e., to retain their optimality across several changes without requiring re-tuning. We now evaluate if a recipe tailored for one code version is also tailored for another, even when the two versions are 50 commits apart.

For this experiment, we took a random sample of 60 files from our benchmarks and retrieved the 50 most recent commits per file. We only sampled 60 out of 120 files as building these files for each commit is quite time consuming—it can take up to a couple of days. We instrumented each file version with the four assertion types described in Sect. V-B. It should be noted that, for some files, we retrieved fewer than 50 versions either because there were fewer than 50 total commits or our build procedure for the project failed on older commits. This is also why we did not run this experiment for over 50 commits.

We analyzed each file version with the best recipe, R_o, found by TAILOR for the oldest file version. We compared this recipe with new best recipes, R_n, that were generated by TAILOR when run on each subsequent file version. For this experiment, we used a 5-min timeout and 40 iterations.

Note that, when running TAILOR with the same optimization algorithm and random seed, it explores the same recipes. It is, therefore, very likely that recipe R_o for the oldest commit is also the best for other file versions since we only explore 40 different recipes. To avoid any such bias, we performed this experiment by seeding TAILOR with a different random seed for each commit. The results are shown in Figs. 12 and 13.

In Fig. 12, we give a bar chart comparing the number of files per commit that have a positive, equal, and negative difference in the number of verified assertions, where commit 0 is the oldest commit and 49 the newest. An equal difference (in orange) means that recipe R_o for the oldest commit proves the same number of assertions as the current file version, f_n, as recipe R_n found by running TAILOR on f_n. To be more precise, we consider the two recipes to be equal if they differ by at most 1 verified assertion or 1% of verified assertions since such a small change in the number of safe assertions seems acceptable in practice (especially given that the total number of assertions may change across commits). A positive difference (in green) means that R_o achieves better verification results than R_n, that is, R_o proves more assertions safe (over 1 assertion or 1% of the assertions that R_n proves). Analogously, a negative difference (in red) means that R_o proves fewer assertions. We do not consider time here because none of the recipes timed out when applied on any file version.

Note that the number of files decreases for newer commits. This is because not all files go forward by 50 commits, and even if they do, not all file versions build. However, in a few instances, the number of files increases going forward in time. This happens for files that change names, and later, change back, which we do not catch.

For the vast majority of files, using recipe R_o (found for the oldest commit) is as effective as using R_n (found for the current commit). The difference in safe assertions is negative for less than a quarter of the files tested, with the average negative difference among these files being around 22% (i.e., R_o proved 22% fewer assertions than R_n in these files). On the remaining three quarters of the files tested however, R_o proves at least as many assertions as R_n, and thus, R_o tends to be tailored across code versions.

Fig. 13 shows the average difference in the number of verified assertions per change in lines of code from the oldest commit. Note that a positive (resp. negative) difference represents that R_o (resp. R_n) proves more assertions. The plot clearly shows that for most files, regardless of lines changed, R_o and R_n are equally effective. Regarding the outliers shown in the figure, we noticed that all commits with a difference of 50 safe assertions or more modify one
particular file from the Git project. In this case, $R_o$ is not as effective because the widening and narrowing settings have very low values.

**RQ4 takeaway:** For over 75% of files, TAILOR’s recipe for a previous commit (from up to 50 commits previous) remains tailored for future versions of the file, indicating the resilience of tailored recipes across code changes.

### D. Threats to Validity

We have identified the following threats to the validity of our experiments.

**Benchmark selection.** Our results may not generalize to other benchmarks. However, we selected popular GitHub projects from different application domains (see Tab. 2). Hence, we believe that our benchmark selection mitigates this threat and increases generalizability of our findings.

**Abstract interpreter and recipe settings.** For our experiments, we only used a single abstract interpreter, CRAB, which however is a mature tool and actively supported. The selection of recipe settings was, of course, influenced by the available settings in CRAB. Nevertheless, CRAB implements the generic architecture of Fig. 2, used by most abstract interpreters, such as those mentioned at the beginning of Sect. III. We, therefore, expect our approach to generalize to such analyzers.

**Optimization algorithms.** We considered four optimization algorithms, but in Sect. IV-C, we explain why these are suitable for our application domain. Moreover, TAILOR is configurable with respect to the optimization algorithm.

**Assertion types.** Our results are based on four types of assertions. However, these cover a broad spectrum of runtime errors that are commonly checked by static analyzers.

### VI. Related Work

The impact of different abstract-interpretation configurations has been previously evaluated [47] for Java programs and partially inspired this work. To the best of our knowledge, we are the first to propose tailoring static analyzers to custom usage scenarios using optimization. However, optimization is a widely used technique in many engineering disciplines. In the following, we focus on its use in program analysis.

Optimization has been successfully applied to a number of program-analysis problems, such as automated testing [43], [44], invariant inference [45], and compiler optimizations [46].

Many machine-learning techniques also rely on optimization, for instance of loss functions in neural networks. Recently, researchers have started to explore the direction of enriching program analyses with machine-learning techniques, for example, to automatically learn analysis heuristics [49], [50], [51], [52]. A particularly relevant body of work is on adaptive program analysis [54], [55], [56], where existing code is analyzed to learn heuristics that trade soundness for precision or that coarsen the analysis abstractions to improve memory consumption.

More specifically, adaptive program analysis poses different static-analysis problems as machine-learning problems and relies on Bayesian optimization to solve them, e.g., the problem of selectively applying unsoundness to different program components (e.g., different loops in the program) [56]. The main insight is that program components (e.g., loops) that produce false positives are alike, predictable, and share common properties. After learning to identify such components for existing code, this technique suggests components in unseen code that should be analyzed unsoundly.

In contrast, TAILOR currently does not adjust soundness of the analysis. However, this would also be possible if the analyzer provided the corresponding configurations. More importantly, adaptive analysis focuses on learning analysis heuristics based on existing code in order to generalize to arbitrary, unseen code. TAILOR, on the other hand, aims to tune the analyzer configuration to a custom usage scenario, including a particular program under analysis. In addition, the custom usage scenario imposes user-specific resource constraints, for instance by limiting the time according to a phase of the software-engineering life cycle. As we show in our experiments, the tuned configuration remains tailored to several versions of the analyzed program. In fact, it outperforms configurations that are meant to generalize to arbitrary programs, such as the default recipe.

### VII. Conclusion

In this paper, we have proposed a technique and framework that tailors a generic abstract interpreter to custom usage scenarios. We instantiated our framework with a mature abstract interpreter to perform an extensive evaluation on real-world benchmarks. Our experiments show that the configurations generated by TAILOR are vastly better than the default options, vary significantly depending on the code under analysis, and most remain tailored to several subsequent code versions. In the future, we plan to explore the challenges that an inter-procedural analysis would pose, for instance, by using a different recipe for computing a summary of each function or each calling context.
