Contents

Chapter 1

Evaluation

Section **??** described the implementation of the Nickel Language Server addressing the first research question stated in sec. **??**. Proving the viability of the result and answering the second research question demands an evaluation of different factors.

Earlier, the most important metrics of interest were identified as:

Usability What is the real-world value of the language server?
Does it improve the experience of developers using Nickel? NLS offers several features, that are intended to help developers using the language. The evaluation should assess whether the server does improve the experience of developers.
Does NLS meet its users' expectations in terms of completeness, correctness and behavior? Labeling NLS as a Language Server, invokes certain expectations built up by previous experience with other languages and language servers. Here, the evaluation should show whether NLS lives up to the expectations of its users.

Performance What are the typical latencies of standard tasks? In this context latency refers to the time it takes from issuing an LSP command to the reply of the server. The JSON-RPC protocol used by the LSP is synchronous, i.e. requires the server to return results of commands in the order it received them. Since most commands are sent implicitly, a quick processing is imperative to avoid commands queuing up.
Can single performance bottlenecks be identified? Single commands with excessive runtimes can slow down the entire communication resulting in bad user experience. Identified issues can guide the future work on the server.
How does the performance of NLS scale for bigger projects? With increasing project sizes the work required to process files increases as well. The evaluation should allow estimates of the sustained performance in real-world scenarios.

Answering the questions above, this chapter consists of two main sections. The first section sec. 1.2 introduces methods employed for the evaluation. In particular, it details the survey (sec. 1.4.1) which was conducted with the intent

to gain qualitative opinions by users, as well as the tracing mechanism (sec. 1.2.3) for factual quantitative insights. Section 1.4 summarizes the results of these methods.

1.1 Evaluation Considerations

Different methods to evaluate the abovementioned metrics were considered. While quantifying user experience yields statistically sound insights about the studied subject, it fails to point out specific user needs. Therefore, this work employs a more subjective evaluation based on a standardized experience report focusing on individual features. Contrasting the expectations highlights well executed, immature or missing features. This allows more actionable planning of the future development to meet user expectations.

On the other hand it is more approachable to track runtime performance objectively through time measurements. In fact, runtime behavior was a central assumption underlying the server architecture. As discussed in sec. **??** NLS follows an eager, non-incremental processing model. While incremental implementations are more efficient, as they do not require entire files to be updated, they require explicit language support, i.e., an incremental parser and analysis. Implementing these functions exceeds the scope of this work. Choosing a non-incremental model on the other hand allowed to reuse entire modules of the Nickel language. The analysis itself can be implemented both in a lazy or eager fashion. Lazy analysis implies that the majority of information is resolved only upon request instead of ahead of time. That is, an LSP request is delayed by the analysis before a response is made. Some lazy models also support memoizing requests, avoiding to recompute previously requested values. However, eager approaches preprocess the file ahead of time and store the analysis results such that requests can be handled mostly through value lookups. To fit Nickels' type-checking model and considering that in a typical Nickel workflow, the analysis should still be reasonably efficient, the eager processing model was chosen over a lazy one.

1.2 Methods

1.2.1 Objectives

The qualitative evaluation was conducted with a strong focus on the first metric in [sec:metrics]. Usability proves hard to quantify, as it is tightly connected to subjective perception, expectations and tolerances. The structure of the survey is guided by two additional objectives, endorsing the separation of individual features. On one hand, the survey should inform the future development of NLS; which feature has to be improved, which bugs exist, what do users expect. This data is important for NLS both as an LSP implementation for Nickel (affecting the perceived maturity of Nickel) and a generic basis for other projects. On the other hand, since all features are essentially queries to the common linearization data structure (cf. [sec. **??**), the implementation of this central structure is an essential consideration. The survey should therefore also uncover apparent problems with this architecture. This entails the use of language abstractions (cf.

sec. **??**) and the integration of Nickel core functions such as the type checking procedure.

The quantitative study in contrast focuses on measurable performance. Similarly to the survey-based evaluation, the quantitative study should reveal insight for different features and tasks separately. The focus lies on uncovering potential spikes in latencies, and making empirical observations about the influence of Nickel file sizes.

1.2.2 Qualitative Evaluation Setup

Inspired by the work of Leimeister in (**leimeister?**), a survey aims to provide practical insights into the experience of future users. In order to get a clear picture of the users' needs and expectations independently of the experience, the survey consists of two parts – a pre-evaluation and final survey.

1.2.2.1 Pre-Evaluation

1.2.2.1.1 Expected features The pre-evaluation introduced participants in brief to the concept of language servers and asked them to write down their understanding of several LSP features. In total, six features were surveyed corresponding to the implementation as outlined in sec. **??**, namely:

1.2.2.1.2 Expected behaviour

1. Code completion Suggest identifiers, methods or values at the cursor position.
2. Hover information Present additional information about an item under the cursor, i.e., types, contracts and documentation.
3. Jump to definition Find and jump to the definition of a local variable or identifier.
4. Find references List all usages of a defined variable.
5. Workspace symbols List all variables in a workspace or document.
6. Diagnostics Analyze source code, i.e., parse and type check and notify the LSP Client if errors arise. The item for the "Hover" feature for instance reads as follows:

   > Editors can show some additional information about code under the cursor. The selection, kind, and formatting of that information is left to the Language Server.
   >
   > What kind of information do you expect to see when hovering code? Does the position or kind of element matter? If so, how?

Items first introduce a feature on a high level followed by asking the participant to describe their ideal implementation of the feature.

1.2.2.2 Experience Survey

For the final survey, interested participants at Tweag were invited to a workshop introducing Nickel. The workshop allowed participants unfamiliar with the Nickel language to use the language and experience NLS in a more natural setting.

Following the workshop, participants filled in a second survey which focused on three main aspects:

First, the general experience of every individual feature. Without weighing their in expectations, the participants were asked to give a short statement of their experience. The item consists of a loose list of statements with the aim to achieve a rough quality classification:

☐ The feature did not work at all
☐ The feature behaved unexpectedly
☐ The feature did not work in all cases
☐ The feature worked without an issue
☐ Other

The following items survey the perceived performance and stability. The items were implemented as linear scales that span from "Very slow response" to "Very quick response" and "Never Crashed" to "Always Crashed" respectively. The second category asked participants to explicitly reflect on their expectations:

○ The feature did not work at all
○ Little of my expectation was met
○ Some expectations were mete, enough to keep using NLS for this feature
○ Most to all expectations were met ○ NLS surpassed the expectations
○ Other

In the final part participants could elaborate on their answers.

Why were they (not) satisfied?
What is missing, what did they not expect?

1.2.3 Quantitative

To address the performance metrics introduced in sec. **??**, a quantitative study was conducted, that analyzes latencies in the LSP-Server-Client communication. The study complements the subjective reports collected through the survey (cf. sec. 1.2.2.2). The evaluation is possible due to the inclusion of a custom tracing module in NLS. The tracing module is used to create a report for every request, containing the processing time and a measure of the size of the analyzed document. If enabled, NLS records an incoming request with an identifier and time stamp. While processing the request, it adds additional data to the record, i.e., the type of request, the size of the linearization (cf. sec. **??**) or processed file and possible errors that occured during the process. Once the server replies to a request, it records the total response time and writes the entire record to an external file.

The tracing approach narrows the focus of the performance evaluation to the time spent by NLS. Consequently, the performance evaluation is independent of the LSP client (editor) that is used. Unlike differences in hardware which affects all operations similarly, LSP clients may implement different behaviors that may cause editor-specific biases. For instance, the LSP does not specify the frequency at which file changes are detected, which in turn can lead to request queuing depending on the editor used.

1.3 Process

1.4 Results

1.4.1 Qualitative

As outlined in [#sec:qualitative-study-outline], the qualitative study consists of two parts conducted before and after an introductory workshop. The pre-evaluation aimed to catch the users' expected features and behaviors, while the main survey asked users about their concrete experiences with the NLS.

1.4.1.1 Pre-Evaluation

In the initial free assessment of expected features (c.f. [#sec:expected-features]) the participants unanimously identified four of the six language server capabilities that guided the implementation of the project (c.f. sec. **??**): Type-information on hover, automactic diagnostics, Code Completion and Jump-to-Definition.

The other two features, Find-References and Workspace/Document Symbols on the contrary were sparingly commented. Some participands noted that they did not use these capabilities.

1.4.1.1.1 Type-information on hover Hovering is expected to work on values as well as functions. For values it is desired to show types including applied contracts, documentation and defautl values. On functions it should display the function's signature and documentation. Additionally hovering an item desireably visualizes the scope of the item, i.e. where it is available.

1.4.1.1.2 Diagnostics Diagnostics are expected to include error messages signalling syntax and type errors as well as possibly evaluation errors and contract breaches. The diagnostics should show up at the correct positions in the code and "suggest how to fix" mistakes. Code linting was named as a possible extension to error reporting. This would include warnings about bad code style – formatting, casing conventions – unused variables, deprecated code and undocumented elements. Moreover structural analysis was conceived to allow finding structural issues and help fixing them In either case the diagnostic should be produced "On-the-fly" while typing or upon saving the document.

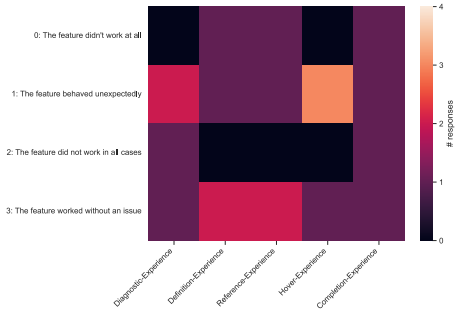
1.4.1.1.3 Code Completion Code Completion was described as a way to chose from possible completion candidates of options. Completable items can be variable names, record fields, types or functions. Besides, Participants conceived filtereing or prioritizing of candidates by type if applied as function arguments. Finally, the completion context could guide priorization as well as auto-generation of contract and function skeletons.

1.4.1.1.4 Jump-to-Definition Users expect Jump-to-Definition to work with any kind of reference i.e., variable usages, function calls, function arguments and type annotations. On records and references to records, users expect statically defined nested fields to point to the correct respective definition. The ability to define self referencing records was however conceded to be a challenge.

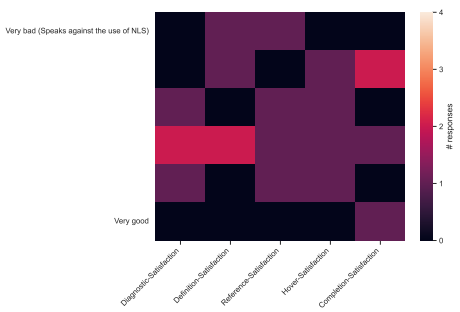
1.4.1.1.5 Other features Syntax highlighting and code formatting as well as error tolerance were named as further desireable features of a language server beyond the explicitly targeted features. Error tolerance was detailed as the capability of the language server to continue processing and delivering analysis of invalid sources. For invalid files a language server should still be able to provide its functionality for the correct parts of the program.

1.4.1.2 Experience Survey

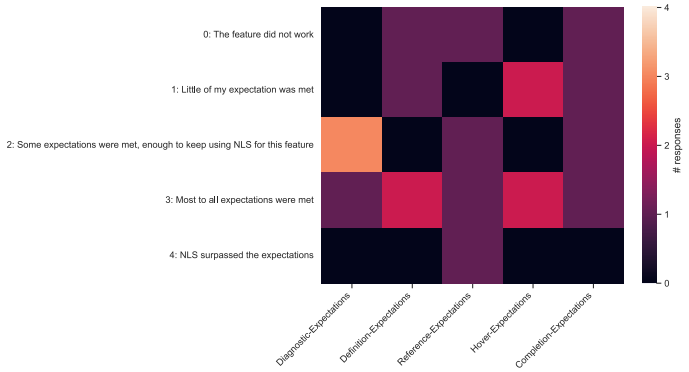
This subsection describes the results from the filled after the Nickel workshop in which participants were asked to install the LSP to support their experience. It first looks at a summary of the data, before diving into the comments for each directly addressed feature.



(a) Number of votes for each experience category and feature



(b) Number of votes for each satisfaction category and feature



(c) Number of votes for each expectation category and feature

Figure 1.1: User responses regarding general experience, fulfillment of expectations and general satisfaction.

The above figures show the turnout of three items from the survey for each of the relevant features. Neither of them shows clear trends with positive and negative results distributed almost evenly between positive and negative sentiment.

The first graph (fig. 1.1a) represents the participants' general experience with the relevant features. It shows that each feature worked without issue in at least

one instance. Yet, three features were reported to not work at all and no feature left the users unsurprised. Users found the hover and diagnostic features to behave particularly unexpectedly.

In the second item of each feature, the survey asked the subjects to rate the quality of the language server based on their expectations. Figure 1.1c summarizes the results. In agreement with the first graph, one user was unable to use at least three features entirely. The majority of responses show that NLS met its user's expectations at least partially. The results are however highly polarized as the Jump-to-Definition and Hover features demonstrate; Each received equally many votes for being inapt and fully able to hold up to the participants expectations at the same time. Other features were left with a uniformly distributed assessment (e.g. Completion and Find-References). The clearest result was achieved by the Diagnostics feature, which received a slight but uncontested positive sentiment.

The general satisfaction with each feature was answered in the same polarized manner as seen in fig. 1.1b. A slight majority of responses falls into the upper half of the possible spectrum. Two of the features reported without function in the preceding questions were given the lowest possible rating.

1.4.1.2.1 Hover {#sec:hover@res} As apparent in (fig. 1.1a), most participants experienced unexpected behavior by the LSP when using the hover functionality. In the comments, extraneous debug output and incorrect displaying of the output by the IDE are pointed out as concrete examples. However, one answer suggests that the feature was working with "usually useful" output.

1.4.1.2.2 Diagnostics {#sec:diagnostics@res} While the diagnostics shown by NLS appear to behave unexpectedly for some users in fig. 1.1a, no user felt deterred from keep using NLS for it as displayed in fig. 1.1c. Some respondents praised the "quick" and "direct feedback" as well as the visual error markers pointing to the exact locations of possible issues. On the contrary, others mentioned "unclear messages" and pointed out that contracts were not checked by the Language Server. Moreover, a performance issue was brought up noting that in some situations NLS "queues a lot of work and does not respond."

1.4.1.2.3 Code Completion {#sec:code-completion@res} Comments about the Code Completion feature were unanimously critical. Some participants noted the little gained "value over the token based completion built into the editor" while others specifically pointed at "missing type information and docs." Additionally, record field completion was found to be missing, albeit highly valued.

1.4.1.2.4 Document Navigation {#sec:document-navigation@res} Results and comments about the Go-To-Definition and Find-References were polarized. Some users experienced unexpected behavior or were unable to use the feature at all (cf. fig. 1.1a). Similarly, the comments on one hand suggest that "the feature works well and is quick" while on the other mention inconsistencies and unavailability. More specifically, cross file navigation was named an important missing feature.

1.4.1.2.5 General Performance {#sec:general-performance@res} The responses to the general performance suggest that NLS' performance is largely dependent on its usage. On unmodified files queries were reported to evaluate "instantaneously." However, modifying files caused that "modifications stack up" causing high CPU usage and generally "very slow" responses. Besides, documentation was reported as slow to resolve while the server itself was "generally fast."

1.4.2 Quantitative

The quantitative evaluation focuses on the performance characteristics of NLS. As described in sec. **??** a tracing module was embedded into the NLS binary which recorded the runtime together with the size of the analyzed data, i.e., the number of linearization items sec. **??** or size of the analyzed file. This section will first introduce the dataset before looking at the general performance and finally looking into particular cases.

1.4.2.1 Dataset

The underlying data set consists of 16760 unique trace records. Since the `textDocument/didOpen` method is executed on every update of the source, it greatly outnumbers the other events. The final distribution of methods traced is:

Table 1.1: Number of traces per LSP method

Method	count	linearization based
`textDocument/didOpen`	13436	no
`textDocument/completion`	2981	yes
`textDocument/hover`	227	yes
`textDocument/definition`	68	yes
`textDocument/references`	49	yes
total	16761	

Figures 1.2 break up these numbers by method and linearization size or file size respectively. The linearization is the linear representation of an enriched AST. It is explained in great detail in sec. **??**. The first figure shows a peak number of traces for completion events between 0 to 1 linearization items as well as local maxima around a linearization size of 20 to 30 and sustained usage of completion requests in files of $90 - 400$ items. Similar to the completion requests (but well outnumbered in total counts), other methods were used mainly in the range between 200 and 400 linearization items. A visualization of the Empirical Cumulative Distribution Function (ECFD) fig. **??** corroborates these findings. Moreover, it shows an additional hike of Jump-to-Definition and Find-References calls at on files with around 1500 linearization items. The findings for linearization based methods line up with those depicting linearization events (identified as `textDocument/didOpen`). An initial peak referring to rather small input files between 300 and 400 bytes in size is followed by a sustained usage of the NLS on files with 2 to 6 kilobytes of content topped with a final application

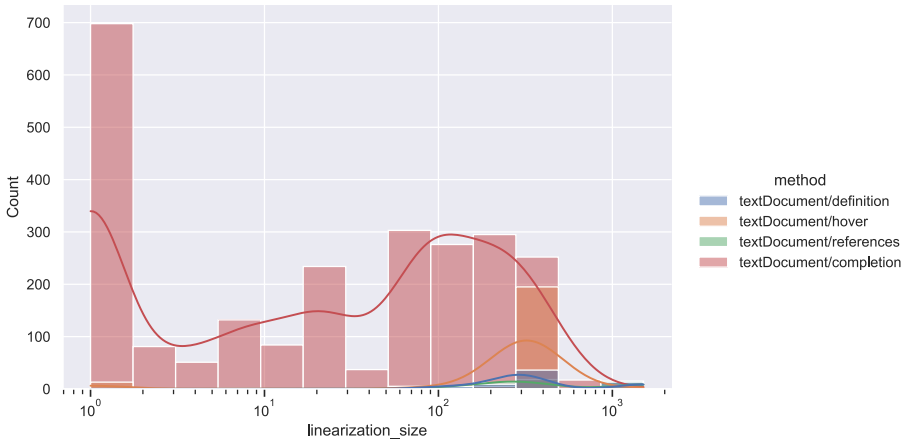


Figure 1.2: Distribution of linearization based LSP requests

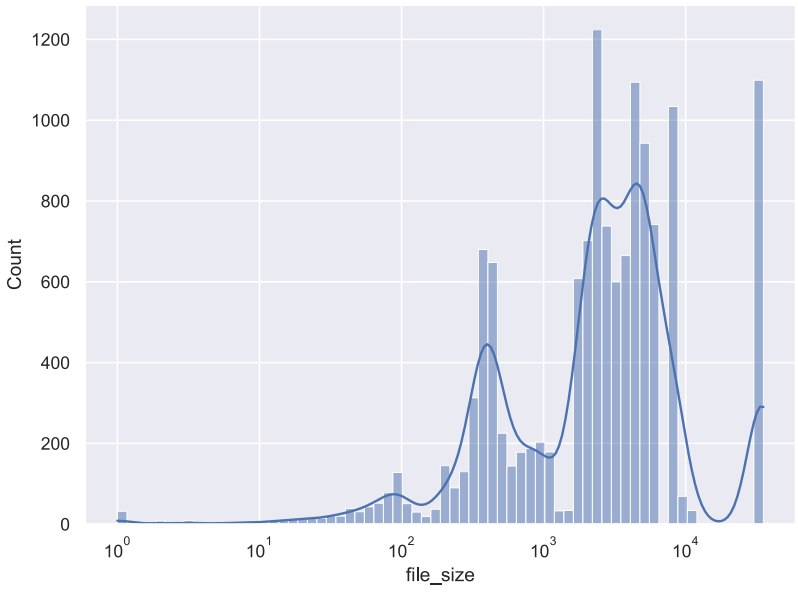


Figure 1.3: Distribution of file analysis requests

on 35 kilobyte large data.

1.4.2.2 Big Picture Latencies

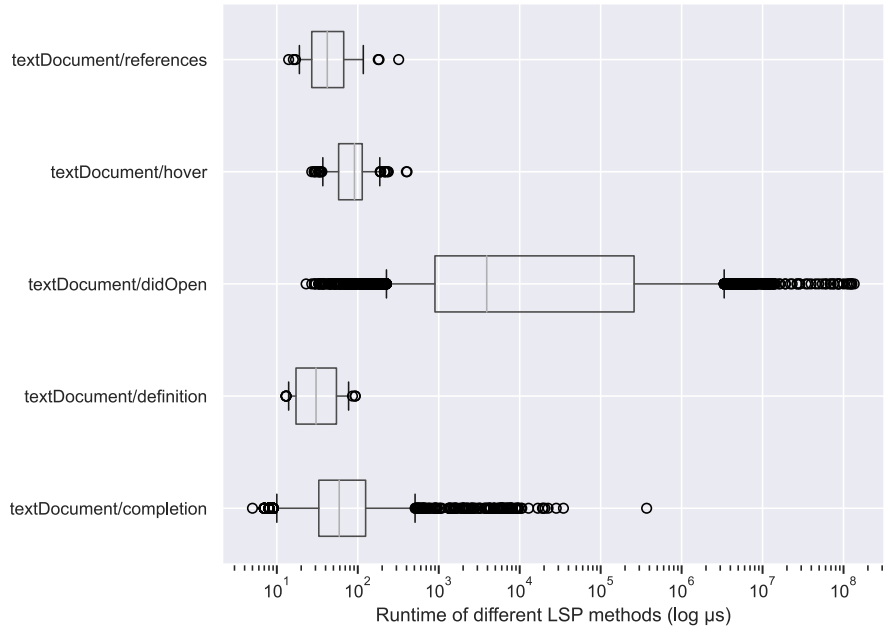
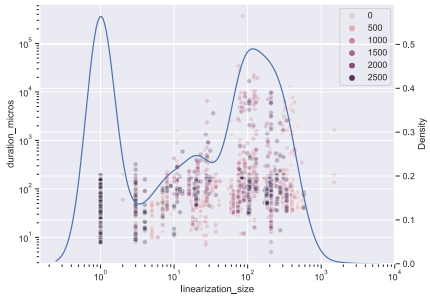


Figure 1.4: Statistical runtime of different LSP methods

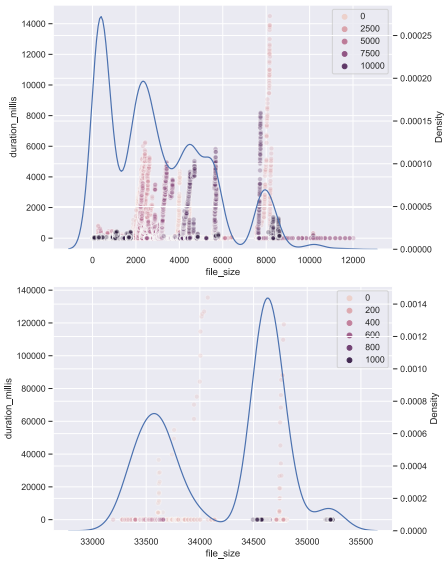
Comparing the runtime of the individual methods alone in fig. 1.4, reveals three key findings. First, all linearization based methods exhibit a sub-millisecond latency in at least 95 of all invocations and median response times of less than $100ţs$ However, maximum latencies of completion invocations reached tens of milliseconds and in one recorded case about $300ms$. Finally, document linearization as associated with the `textDocument/didOpen` method shows a great range with maxima of $1.5 * 10^5 ţs$ (about 2.5 minutes) and a generally greater interquartile range spanning more than two orders of magnitude.

1.4.2.3 Special cases

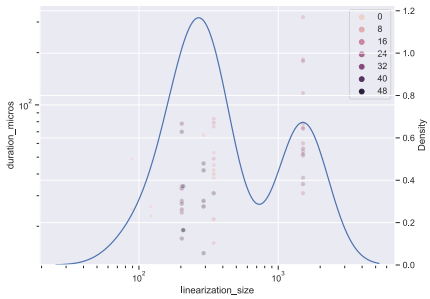
Setting the runtime of completion requests in relation to the linearization size on which the command was performed shows no clear correlation between the dimensions. In fact the correlation coefficient between both variables measures 0.01617 on a linear scale and 0.26 on a $\log_{10} \log_{10}$ scale. Instead, vertical columns stand out in the correlation graph fig. 1.5a. The height of these columns varies from one to five orders of magnitude. The item density shows that especially high columns form whenever the server receives a higher load of requests. Additionally, color coding the individual requests by time reveals that the trace points of each column were recorded at a short time interval. Applying the same analysis to the other methods in figs. 1.5b, 1.5c, **??** returns similar findings, although the columns remain more compact in comparison to the Completions method. In case



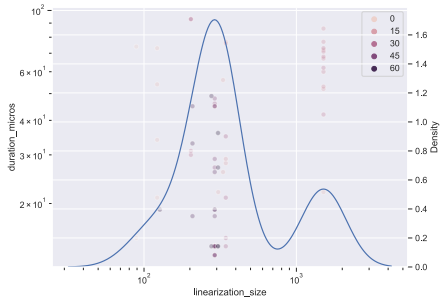
(a) Runtime latencies of completion requests at differnet linearization sizes



(b) Runtime latencies of hover requests at differnet linearization sizes



(c) Runtime latencies of find-references requests at differnet linearization sizes



(d) Runtime latencies of jump-to-definition requests at differnet linearization sizes

Figure 1.5: Runtime latencies of different linearization based methods

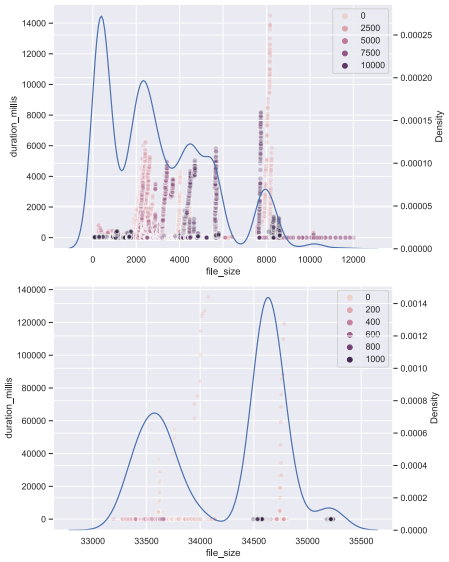


Figure 1.6: Runtime latencies of file update handlings at different file sizes

of the `didOpen` method columns are clearly visible too [#fig:correlation-opens]. However, here they appear leaning as suggesting an increase in computation time as the file grows during a single series of changes to the file.