ABSTRACT

NAIR, VIVEK. Frugal ways to find Good Configurations. (Under the direction of Dr. Timothy J. Menzies.)

Most software systems available today are configurable, which gives the users an option to customize the system to achieve differ- ent functional or non-functional (better performance) properties. As systems evolve, more configuration options are added to the software system. Studies report that developers find it difficult to understand the configuration spaces, which leaves considerable optimization potential untapped and induces major economic cost. To solve this problem of finding the (near) optimal configurations, engineers have proposed various techniques. Most popular among them are model-based techniques, where accurate models of the configuration space are created using as few configuration measurements as possible. We notice two major problems with the model-based techniques: 1) previous techniques are expensive to be practically viable, and 2) there are software systems whose configuration spaces cannot be accurately modeled. Consequently, there is a gap between proposed techniques and practical viability of these techniques. This dissertation will focus on proposing techniques which are easier to understand and is practically viable.

First, we present **WHAT** that exploits some lower dimensional knowledge to build performance models. Prior work on predicting the performance of software configurations suffered from either (a) requiring far too many sample configurations or (b) large variances in their predictions. Both these problems can be avoided using the **WHAT** spectral learner. WHAT's innovation is the use of the spectrum (eigenvalues) of the distance matrix between the configurations of a configurable software system, to perform dimensionality reduction. Within that reduced configuration space, many closely associated configurations can be studied by executing only a few sample configurations. For the subject systems studied here, a few dozen samples yield accurate and stable predictors?less than 10 % prediction error, with a standard deviation of less than 2%. When compared to the state of the art, WHAT (a) requires 2 to 10 times fewer samples to achieve similar prediction accuracies, and (b) its predictions are more stable (i.e., have lower standard deviation). Furthermore, we demonstrate that predictive models generated by WHAT can be used by optimizers to discover system configurations that closely approach the optimal performance.

The second contribution is a rank-based method which shows how an accurate model is not required for performance optimization, but a rank-preserving model is sufficient. We evaluate rank-based method with 21 scenarios based on nine software systems and demonstrate that our approach is beneficial in 16 scenarios; for the remaining five scenarios, an accurate model can be built by using very few samples anyway, without the need for a rank-based approach.

Additionally, in 8/21 of the scenarios, the number of measurements required by the rank-based method is an order of magnitude smaller than methods used in prior work. To further improve our second contribution, we also propose a Bayesian-based method called Flash: an alternative to model-based technique. Based on preliminary evidence, which can further reduce the cost of performance optimization.

Finally, we present **FLASH**, a sequential model-based method, which sequentially explores the configuration space by reflecting on the configurations evaluated so far to determine the next best configuration to explore. FLASH scales up to software systems that defeat the prior state of the art model-based methods in this area. FLASH runs much faster than existing methods and can solve both single-objective and multi-objective optimization problems. The central insight of this paper is to use the prior knowledge (gained from prior runs) to choose the next promising configuration. This strategy reduces the effort (ie, number of measurements) required to find the (near) optimal configuration. We evaluate FLASH using 30 scenarios based on 7 software systems to demonstrate that FLASH saves effort in 100% and 80% of cases in single-objective and multi-objective problems respectively by up to several orders of magnitude compared to the state of the art techniques.

Frugal ways to find Good Configurations

by
Vivek Nair

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina

2018

APPROVED BY:

Dr. Kathyrn Stolee	Dr. Min Chi
Dr. Ranga Raju Vatsavai, Ph.D.	Dr. Timothy J. Menzies Chair of Advisory Committee

DEDICATION

To Ammayi—for timely advice and support

BIOGRAPHY

Vivek Nair was born and raised in the city of Kolkata, India. He graduate from West Bengal University of Technology and National Institute of Technology, Durgapur with and Bachelors in Technology and Master in Technology respectively. His primary research interest is search based software engineering with a primary goal to solve search problems to maximize the quality of the search while minimizing the cost of search. His current research investigated problems in the software engineering domain and developed techniques not only to solve software configuration problem, but also software product lines as well as cloud architecture tuning problems. Before joining the PhD program, he had two years of professional experience in the Samsung Software Engineering Labs. During his PhD program, his internship experiences include Lexisnexis Risk Solutions (2015-2017), and Microsoft Research - Redmond, USA (Summer 2018). He is a student member of IEEE.

ACKNOWLEDGEMENTS

**This acknowledgment contains plenty of subtexts and would be found as text within brackets. The subtext is meant to serve the purpose of comic relief and comic relief only.*

This thesis is no way an individual work. It is a product of a lot of love and sacrifice, tears and sweat of not just me but for various individuals who have taught lessons throughout my journey as a graduate student, here at NCSU. This text below is no way a comprehensive list of all the people who have touched my life [for better or for worse], but it has been a useful experience.

One of the primary reasons, I can write this thesis is because of Dr. George N. Rouskas and my aunt, Dr. Latha Unni. My first year at NCSU was tumultuous and eventful. As new graduate student, dealing with the stresses of graduate school and the toxic environment of my previous lab was overwhelming. This resulted in me deciding to quit graduate school—which I have been so excited to attend. Dr. Rouskas, who provided me with financial support while transitioning from my old advisor to my current advisor. Both of them counseled and provided support when I required it the most. This thesis would have never been possible without these two amazing people.

Secondly my thesis advisor [my academic father], Dr. Menzies—he is a legend [faking the Aussie accent]. He has been very [not so] patient with me and directed me towards areas which held promise. He helped me understand the value of collaboration and being a voracious reader of the literature. That said I have never had a more dramatic relationship with anyone—which in retrospect was an effective way of training graduate student. Our story has all the elements of a soap opera—there are love and respect, not so much love [hate sounds just too strong] and doubt. Either way, he played a massive part in making this thesis possible and made me a better researcher [hopefully].

Thirdly, I would then like to thank my dissertation committee members, Dr. Ranga Raju Vatsavai, Dr. Min Chi and Dr. Kathryn T. Stolee, for their valuable feedback and insight on my dissertation.

I am also grateful to my research collaborators: Sven Apel (University of Passau), Norbert Siegmund (Bauhaus-University Weimar), and Pooyan Jamshidi (University of South Carolina). I would like to gratefully acknowledge researchers who generously shared their research tools and results used in my dissertation. I am very much thankful to all my peers at the RAISE Lab, for their constant support, insightful discussions and useful feedback on my research. Dr. Wei Fu, Dr. Chin-Jung Hsu, Rahul Krishna, George Mathew [also my room-mate], and Zhe Yu for the insightful and exciting conversations [interestingly I was able to write papers with all these

amazing individuals—so all our discussions were not completely useless]. I thank my internship mentors in the industry, who have immensely helped me expand my research to a broader scope: Dan Camper and Arjuna Chala (Lexisnexis Risk Solutions), and Chris Duvarney, Kim Herzig, and Hitesh Sajnani (Microsoft Research, USA).

Heartfelt thanks go to Kathy Luca, Carol Allen, and all the helpful staff at the Department of Computer Science. I thank Amruthkiran Hedge, Anand Gorthi, Sandesh Saokar, Siddartha Chauhan, Mayank Vaish, Akhilesh Tanneru [Mr. T], and George Mathew for sharing the apartment and making my stay at Raleigh eventful. Special thanks to Karen Warmbein, Blue and [Z]Simba whose love and support helped me survive the grind of Ph.D.

Finally, last but not least, I am very thankful to my family. My family was my constant source of support and encouragement throughout my journey as a doctoral student. My Ph.D. journey would not have gone so far, without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Modern software systems nowadays provide configuration options to modify both functional behavior of the system, i.e. functionality of the system, and non-functional properties of the system, such as performance and memory consumption. Configuration options of a software system that are relevant to users are usually referred to as features. All the features of a system (vector of configuration options) together defines a *configuration* of a software system. The features can often taken integer, decimal or string values. One of the most important non-functional properties is performance, because it influences the how a user interacts with the system. Performance can be influenced by many factors including the environment (for example, the hardware in which the software system is current executing). A software system is required to select and set configuration options to maximize the performance of that system. For example, say we have a software system with 10 (binary) configuration options—it results in a configuration space of size 2^{10} or 1024. The user of the software system, now has to find the optimal configuration for the given task (or input) in hand. This problem can be tackled in two different ways: (1) exhaustively measuring performance of all possible configurations—which means running 1024 benchmark runs, and (2) use domain knowledge (assuming the user has tuned similar software system before) to find the best configuration. However, as the number of configuration options increase, it becomes difficult for humans to keep track of the interactions between the configuration options. This means as the configuration space grows (size of the configuration space grows exponentially) it is harder to either exhaustively measure performance for all possible configurations or find domain experts to confidently do so. Please note that the optimal configuration we are trying to find can change dramatically with different inputs (tasks) and the environment—which make domain knowledge based decision less reliable.

This exact problem has been reported by numerous researchers from different domains.

- Many software systems have poorly chosen defaults [1], [2]. Hence, it is useful to seek better configurations.

- Understanding the configuration space of software systems with large configuration spaces is challenging [3].
- Exploring more than just a handful of configurations is usually infeasible due to long benchmarking time [4].

The problem we are trying to tackle throughout this document is: "How can be find a set of configuration options which would maximize the performance of a system while minimizing the cost of search". Here were would limit our scope of study to just the configurations options or features of a particular software system (and not its environment).

Conf.	Features																Perf. (s)
$\mathbf{c}_i$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	y_i
$\mathbf{c}_1$	1	1	0	0	1	1	0	0	1	0	0	1	1	1	0	0	292
$\mathbf{c}_2$	1	0	0	0	1	1	0	1	1	1	0	0	1	0	1	0	571
$\mathbf{c}_3$	1	1	1	0	0	1	0	0	1	0	1	0	1	0	0	1	681
$\mathbf{c}_4$	1	1	0	1	1	0	0	0	1	0	1	0	1	1	0	0	263
$\mathbf{c}_5$	1	1	1	0	1	1	0	0	1	0	1	0	1	0	1	0	536
$\mathbf{c}_6$	1	0	0	1	0	1	1	1	1	0	1	0	1	1	0	0	305
$\mathbf{c}_7$	1	1	1	1	0	1	1	0	1	0	1	0	1	0	1	0	408
$\mathbf{c}_8$	1	0	0	1	1	0	1	1	1	0	1	0	1	1	0	0	278
$\mathbf{c}_9$	1	0	0	1	1	0	0	1	1	0	0	1	1	0	1	0	519
$\mathbf{c}_{10}$	1	1	0	0	1	1	0	1	1	0	1	0	1	0	0	1	781
$\mathbf{c}_{11}$	1	0	0	1	1	1	0	1	1	1	0	0	1	0	0	1	822
$\mathbf{c}_{12}$	1	1	0	0	1	0	0	0	1	0	1	0	1	0	0	1	713
$\mathbf{c}_{13}$	1	0	1	0	0	0	1	0	1	0	0	1	1	0	1	0	381
$\mathbf{c}_{14}$	1	0	1	1	1	1	0	1	1	1	0	0	1	0	1	0	564
$\mathbf{c}_{15}$	1	1	1	1	1	0	0	0	1	0	0	1	1	0	1	0	489
$\mathbf{c}_{16}$	1	1	1	1	0	1	1	1	1	1	0	0	1	1	0	0	275

Figure 1.1 A sample of 16 randomly-selected configurations of x264 and corresponding performance measurements (seconds)

1.1 Motivating Example

To motivate our work, we use the same example from a previous work [15], a configurable command-line tool x264 for encoding video streams into the H.264/MPEG-4 AVC format. In this example, we consider 16 encoder features of x264, such as encoding with multiple reference frames and parallel encoding on multiple CPUs. The user can select different features to encode a video. The encoding time is used to indicate the performance of x264 in different configurations. A configuration represents a program variant with a certain selection of features. This example with only 16 features gives rise to 1,152 configurations. Intuitively, 16 binary features should

provide 216 different configurations, however, in this work we consider only valid configurations i.e. configurations that are allowed by the system under investigation. In practice, often only a limited set of configurations can be measured, either by simulation or by monitoring in the field. For example, Figure 1.1 lists a sample of 16 randomly selected configurations and their actual performance measurements. How can we determine the performance of other configurations based on a small random sample of measured configurations? To formulate the above issue, we represent a feature as a binary decision variable x (please note that decision variable could be decimal as well). If a feature is selected in a configuration, then the corresponding variable is set to 1, and 0 otherwise. Assume that there are N features in total, all features of a program are represented as a set $X = x_1, x_2, ..., x_N$. A configuration is an N-tuple c, assigning 1 or 0 to each variable. For example, each configuration of x264 is represented by 16-tuple, e.g. c_1 = ($x_1 = 1$, $x_2 = 1$, $x_3 = 0$, $x_4 = 0$, ..., $x_{16} = 1$). All valid configurations of a program are denoted by C. Each configuration c of a program has an actual measured performance value y. Performance values are taken from publicly available dataset deployed with SPLConqueror tool. All performance values of all configurations C form set Y . Suppose that we acquire a random sample of configurations $C_S \in C$ and their actual measured performance values $Y_S \in Y$, together forming sample S. The problem of variability-aware performance prediction is to predict the performance of other configurations in C CS based on the measured sample S. We regard all variables in X as predictors and a configuration's actual performance value y as the response. In other words, we predict a quantitative response y based on a set of categorical predictors X, which is a typical regression problem. *Due to feature interactions, the above issue is reduced to a non-linear regression problem, where the response depends non-linearly on one or more predictors.*

1.2 Dual Axis: Quality and Cost

Quality: Quality of an approach refers to the distance between the solutions returned by the techniques proposed and the ground truth. Quality can be calculated as the rank difference which is the distance between best solution (also referred to as configuration) found by (proposed) techniques to the actual best solution.

Cost: Cost of an approach refers to the effort required by an approach to find a good configurations. In this thesis, we use number of measurements (or benchmark runs) as an proxy to the search effort.

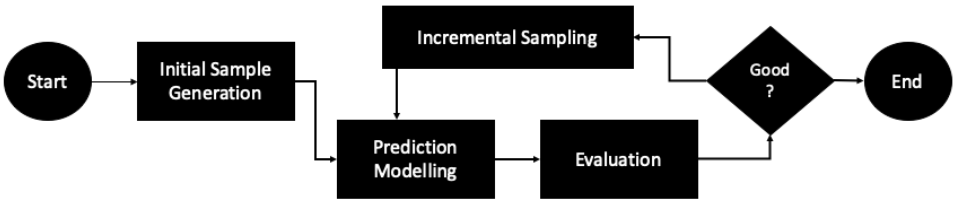

Figure 1.2 General process of performance prediction by sampling

1.3 Prior Work

Figure 1.2 illustrates the general process of performance prediction by sampling. It starts with an initial sample of measured configurations, which are used to build the prediction model. A good initial sample significantly reduces the iterations of the entire prediction process. State-of-the-art approaches fix the size of the initial sample to the number of features or potential feature interactions of a system [15, 39]. However, such a strategy might not be the optimal one, as the number of features (and their interactions) can be high and, at the same time, an acceptable prediction accuracy might be achieved using a substantially smaller set of measured configurations.

Observation: Even though, the strategies proposed by earlier work in this domain exploited random sampling and CART decision trees (which are both scalable), there is repeated work. In our two axis' of evaluation, these strategies work well in the quality aspect however, does not fare well in the cost aspect. Most of the method use over 50% of the configurations (from the configuration space), to build a reliable model.

1.4 Contributions

1.4.1 Clustering

The prior work in this area [15, 33] used a combination of random sampling and regression trees. However, random sampling completely disregards the presence of clusters in the configuration space. It has been shown in the literature [30] that (1) most of the configuration options does not influence the performance i.e. only a few configurations options are useful, and (2) the performance curve is generally a step function.

To elaborate, if we sort the configurations from worst-performance to best and plot configurations along the X-axis and performance along the Y-axis. We expected a continuous graph such as Figure 1.3(a), where high-valued ($) is bad (worst performance is at the far left) and low-valued ($) is good (best performance is at the far right). anchors the far-right point on

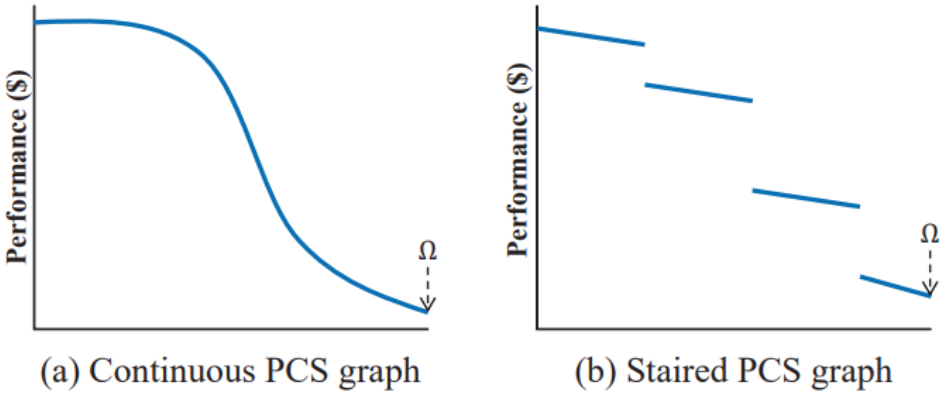

Figure 1.3 Performance curves. (From [30])

X-axis of graphs. Interestingly, Marker et al. [25] discovered that performance curves often occurs (in real world) as stairs, as in Figure 3b. Stairs arise from discrete feature decisions; some features are highly-influential in performance while others have little or no impact. Consequently, a few critical features influences the performance while less important feature decisions alter the performance of nearby configurations only slightly (giving a stair its width and slope).

Intuition: From the literature, it is evident that most of the configuration options in a (given) software system does not affect the performance of the configurations. Hence, the central insight of this work is that random sampling with no regards to this specific feature of this problem is ineffective and hence adds additional cost.

Proposal: This problem can be reformulated as a clustering problem, where we try to find an unsupervised method to cluster the configuration space into meaningful clusters. Once we have this cluster (top-down bi-cluster), we can only use random sampling from each of the clusters. This way, we can reduce redundant measurements (measurements which do not provide new information).

1.4.2 Ranking

Prior work in this area (including [28]), tried to build accurate performance predictors, which (once trained) can be used to predict the performance of a certain configuration. This work reflects on the prior work and asks: "*Our goal is to find a good configuration* [1] *but, why does the prior work transform this problem into building an accurate model?*" Another reason for this question is the very nature of the model building process (previously described in Figure 1.2). We ask the question "How does a user define *Good*"? There is no way for the user to know whether a model can be build with MMRE (Mean Magnitude of Relative Error) less that 10%.

[1] Good is defined as the distance from the optimal configuration

Considering the above mentioned questions, we drastically modify our approach to this problem. We hypothesize that to find the best performing configuration, we do not want a model which can return a predicted performance score which is as close to the actual performance score. Instead, we can build a model which preserves the relative ordering of the configurations.
Intuition: The central insight of this work is that exact performance values (e.g., the response time of a software system) are not required to rank configurations and to identify the optimal one. To elaborate more, let us assume that we have two humans (Adam—134cm, Billy—173cm) (analogous to configurations) and our objective is to identify the tallest person (Billy). To identify the tallest person, do we need a model which accurately predicts their height in a nano-meter scale? We can easily identify Billy even if the bad model predicted Adams height as 700cm and 890cm.
Proposal: We show that, if we (slightly) relax the question we ask, we can build useful predictors using very small sample sets. Specifically, instead of asking "How long will this configuration run?", we ask instead "Will this configuration run faster than that configuration?" or "Which is the fastest configuration?".

1.4.3 Sequential-Model Based Sampling

Prior work in this area primarily used two strategies. Firstly, researchers used machine learning to model the configuration space. The model is built sequentially, where new configurations are sampled randomly, and the quality or accuracy of the model is measured using a holdout set. The size of the holdout set in some cases could be up to 20% of the configuration space [29] and needs to be evaluated (i.e., measured) before even the model is fully built. This strategy makes these methods not suitable in a practical setting since the generated holdout set can be (very) expensive. Secondly, the sequential model-based techniques used in prior work relied on Gaussian Process Models (GPM) to reflect on the configurations explored (or evaluated) so far [50]. However, GPMs do not scale well for software systems with more than a dozen configuration options [44].
Intuition: To reduce the cost of sampling and eliminate the need for holdout set, we use sequential Model-based Optimization (SMBO). SMBO uses the Bayesian methodology to the iterative optimizer by incorporating a prior model (built using configuration which are already measured) on the space of possible target functions, f. By updating this model every time time a configuration is evaluated, a SMBO routine keeps a posterior model of the target function f. This posterior model is the surrogate $f*$ for the function f (ground truth). Figure 1.4 encapsulates the process.
Proposal: We use the intuition present above to develop a technique called FLASH. The key idea of FLASH is to build a performance model that is just accurate enough for differentiating

Algorithm 1 Sequential Model-Based Optimization

Input: $f, \mathcal{X}, S, \mathcal{M}$
$\mathcal{D} \leftarrow \textsc{InitSamples}(f, \mathcal{X})$
for $i \leftarrow |\mathcal{D}|$ **to** T **do**
  $p(y \mid \mathbf{x}, \mathcal{D}) \leftarrow \textsc{FitModel}(\mathcal{M}, \mathcal{D})$
  $\mathbf{x}_i \leftarrow \arg\max_{\mathbf{x} \in \mathcal{X}} S(\mathbf{x}, p(y \mid \mathbf{x}, \mathcal{D}))$
  $y_i \leftarrow f(\mathbf{x}_i)$ ▷ Expensive step
  $\mathcal{D} \leftarrow \mathcal{D} \cup (\mathbf{x}_i, y_i)$
end for

Figure 1.4 Sequential Model-based Optimization. (From http://tiny.cc/3hy80y)

better configurations from the rest of the configuration space. Tolerating the inaccuracy of the model is useful to reduce the cost (measured in terms of the number of configurations evaluated) and the time required to find the better configuration. To increase the scalability of methods using GPM (Gaussian Process Models)—used widely in the machine learning domain, FLASH replaces the GPMs with a fast and scalable decision tree learner.

This dissertation is organized as follows. Chapter 2 presents the background and related work. Next, Chapter 3 describes the design and implementation of *WHAT*, an approach to use lower dimensions to sample configurations to build performance models to achieve accurate and robust performance prediction. Chapter 4 proposes using ranking models instead of regression models to save cost while finding good configurations. In Chapter 5, we design and implement FLASH , an efficient and effective solution to software configuration problem. Finally, Chapter 6 concludes our discussion of software configuration tuning and describes its future directions.

Chapter 2

Background and Related Work

This chapter describes the necessary background that is related to the research problems addressed in this dissertation. We first describe data-intensive computing and its storage architecture. Next we discuss performance prediction for distributed systems. Last, we discuss related work that uses the data-driven approach to optimize system performance.

Chapter 3

Faster Discovery of Faster System Configurations with Spectral Learning

This chapter originally appeared as Nair, V., Menzies, T., Siegmund, N., & Apel, S. (2017). Faster discovery of faster system configurations with spectral learning. Automated Software Engineering, 1-31.

3.1 Abstract

Despite the huge spread and economical importance of configurable software systems, there is unsatisfactory support in utilizing the full potential of these systems with respect to finding performance-optimal configurations. Prior work on predicting the performance of software configurations suffered from either (a) requiring far too many sample configurations or (b) large variances in their predictions. Both these problems can be avoided using the **WHAT** spectral learner. **WHAT** 's innovation is the use of the spectrum (eigenvalues) of the distance matrix between the configurations of a configurable software system, to perform dimensionality reduction. Within that reduced configuration space, many closely associated configurations can be studied by executing only a few sample configurations. For the subject systems studied here, a few dozen samples yield accurate and stable predictors—less than 10 % prediction error, with a standard deviation of less than 2 %. When compared to the state of the art, our approach (a) requires 2 to 10 times fewer samples to achieve similar prediction accuracies, and (b) its predictions are more stable (i.e., have lower standard deviation). Furthermore, we demonstrate that predictive models generated by **WHAT** can be used by optimizers to discover system configurations that closely approach the optimal performance.

3.2 Introduction

Most software systems today are configurable. Despite the undeniable benefits of configurability, large configuration spaces challenge developers, maintainers, and users. In the face of hundreds of configuration options, it is difficult to keep track of the effects of individual configuration options and their mutual interactions. So, predicting the performance of individual system configurations or determining the optimal configuration is often more guess work than engineering. In their recent paper, Xu et al. documented the difficulties developers face with understanding the configuration spaces of their systems [46]. As a result, developers tend to ignore over 5/6ths of the configuration options, which leaves considerable optimization potential untapped and induces major economic cost [46].

Addressing the challenge of performance prediction and optimization in the face of large configuration spaces, researchers have developed a number of approaches that rely on sampling and machine learning [15, 33, 39]. While gaining some ground, state-of-the-art approaches face two problems: (a) they require far too many sample configurations for learning or (b) they are prone to large variances in their predictions. For example, prior work on predicting performance scores using regression-trees had to compile and execute hundreds to thousands of specific system configurations [15]. A more balanced approach by Siegmund et al. is able to learn predictors for configurable systems [39] with low mean errors, but with large variances of prediction accuracy (e.g. in half of the results, the performance predictions for the Apache Web server were up to 50 % wrong). Guo et al. [15] also proposed an incremental method to build a predictor model, which uses incremental random samples with steps equal to the number of configuration options (features) of the system. This approach also suffered from unstable predictions (e.g., predictions had a mean error of up to 22 %, with a standard deviation of up 46 %). Finally, Sarkar et al. [33] proposed a projective-learning approach (using fewer measurements than Guo at al. and Siegmund et al.) to quickly compute the number of sample configurations for learning a stable predictor. However, as we will discuss, after making that prediction, the total number of samples required for learning the predictor is comparatively high (up to hundreds of samples).

The problems of large sample sets and large variances in prediction can be avoided using the **WHAT** spectral learner, which is our main contribution. **WHAT** 's innovation is the use of the spectrum (eigenvalues) of the distance matrix between the configurations of a configurable system, to perform dimensionality reduction. Within that reduced configuration space, many closely associated configurations can be studied by measuring only a few samples. In a number of experiments, we compared **WHAT** against the state-of-the-art approaches of Siegmund et al. [39], Guo et al. [15], and Sarkar et al. [33] by means of six real-world configurable systems: Berkeley DB, the Apache Web server, SQLite, the LLVM compiler, and the x264 video encoder. We found that **WHAT** performs as well or better than prior approaches, while requiring far

fewer samples (just a few dozen). This is significant and most surprising, since some of the systems explored here have up to millions of possible configurations.

Overall, we make the following contributions:

- We present a novel sampling and learning approach for predicting the performance of software configurations in the face of large configuration spaces. The approach is based on a *spectral learner* that uses an approximation to the first principal component of the configuration space to recursively cluster it, relying only on a few points as representatives of each cluster.

- We demonstrate the practicality and generality of our approach by conducting experiments on six real-world configurable software systems (see Figure 3.1). The results show that our approach is more accurate (lower mean error) and more stable (lower standard deviation) than state-of-the-art approaches.

- We report on a comparative analysis of our approach and three state-of-the-art approaches, demonstrating that our approach outperforms previous approaches in terms of sample size and prediction stability. A key finding is the utility of the principal component of a configuration space to find informative samples from a large configuration space.

3.3 Background & Related Work

A configurable software system has a set X of Boolean configuration options,[1] also referred to as features or independent variables in our setting. We denote the number of features of system S as n. The configuration space of S can be represented by a Boolean space Z_2^n, which is denoted by F. All valid configurations of S belong to a set V, which is represented by vectors $\vec{C_i}$ (with $1 \leq i \leq |V|$) in Z_2^n. Each element of a configuration represents a feature, which can either be *True* or *False*, based on whether the feature is selected or not. Each valid instance of a vector (i.e., a configuration) has a corresponding performance score associated to it.

The literature offers two approaches to performance prediction of software configurations: a *maximal sampling* and a *minimal sampling* approach: With *maximal sampling*, we compile all possible configurations and record the associated performance scores. Maximal sampling can be impractically slow. For example, the performance data used in this paper required 26 days of CPU time for measuring (and much longer, if we also count the time required for compiling the code prior to execution). Other researchers have commented that, in real world scenarios, the cost of acquiring the optimal configuration is overly expensive and time consuming [45].

[1] In this paper, we concentrate on Boolean options, as they make up the majority of all options; see Siegmund et al., for how to incorporate numeric options [40].

If collecting performance scores of all configurations is impractical, *minimal sampling* can be used to intelligently select and execute just enough configurations (i.e., samples) to build a predictive model. For example, Zhang et al. [48] approximate the configuration space as a Fourier series, after which they can derive an expression showing how many configurations must be studied to build predictive models with a given error. While a theoretically satisfying result, that approach still needs thousands to hundreds of thousands of executions of sample configurations.

Another set of approaches are the four "additive" *minimal sampling* methods of Siegmund et al. [39]. Their first method, called feature-wise sampling (*FW*), is their basic method. To explain *FW*, we note that, from a configurable software system, it is theoretically possible to enumerate many or all of the valid configurations[2]. Since each configuration ($\vec{C_i}$) is a vector of n Booleans, it is possible to use this information to isolate examples of how much each feature individually contributes to the total run time:

1. Find a pair of configurations $\vec{C_i}$ and $\vec{C_2}$, where $\vec{C_2}$ uses exactly the same features as $\vec{C_i}$, plus one extra feature f_i.

2. Set the run time $\Pi(f_i)$ for feature f_i to be the difference in the performance scores between $\vec{C_2}$ and $\vec{C_i}$.

3. The run time for a new configuration $\vec{C_i}$ (with $1 \leq i \leq |V|$) that has not been sampled before is then the sum of the run time of its features, as determined before:

$$\Pi(C_i) = \sum_{f_j \in C_i} \Pi(f_j) \tag{3.1}$$

When many pairs, such as $\vec{C_1}, \vec{C_2}$, satisfy the criteria of point 1, Siegmund et al. used the pair that covers the *smallest* number of features. Their minimal sampling method, *FW*, compiles and executes only these smallest C_1 and C_2 configurations. Siegmund et al. also offers three extensions to the basic method, which are based on sampling not just the smallest $\vec{C_i}$,$\vec{C_2}$ pairs, but also any configurations with *interactions* between features. All the following minimal sampling policies compile and execute valid configurations selected via one of three heuristics:

PW (pair-wise): For each pair of features, try to find a configuration that contains the pair and has a minimal number of features selected.

HO (higher-order): Select extra configurations, in which three features, f_1, f_2, f_3, are selected if two of the following pair-wise interactions exist: (f_1, f_2) and (f_2, f_3) and (f_1, f_3).

[2]Though, in practice, this can be very difficult. For example, in models like the Linux Kernel such an enumeration is practically impossible [34].

HS (hot-spot): Select extra configurations that contain features that are frequently interacting with other features.

Guo et al. [15] proposed a progressive random sampling approach, which samples in steps of the number of features of the software system in question. They used the sampled configurations to train a regression tree, which is then used to predict the performance scores of other system configurations. The termination criterion of this approach is based on a heuristic, similar to the *PW* heuristics of Siegmund et al.

Sarkar et al. [33] proposed a cost model for predicting the effort (or cost) required to generate an accurate predictive model. The user can use this model to decide whether to go ahead and build the predictive model. This method randomly samples configurations and uses a heuristic based on feature frequencies as termination criterion. The samples are then used to train a regression tree; the accuracy of the model is measured by using a test set (where the size of the training set is equal to size of the test set). One of four projective functions (e.g., exponential) is selected based on how correlated they are to accuracy measures. The projective function is used to approximate the accuracy-measure curve, and the elbow point of the curve is then used as the optimal sample size. Once the optimal size is known, Sarkar et al. uses the approach of Guo et al. to build the actual prediction model.

The advantage of these previous approaches is that, unlike the results of Zhang et al., they require only dozens to hundreds of samples. Also, like our approach, they do not require to enumerate all configurations, which is important for highly configurable software systems. That said, as shown by our experiments (see Section 3.5), these approaches produce estimates with larger mean errors and partially larger variances than our approach. While sometimes the approach by Sarkar et al. results in models with (slightly) lower mean error rates, it still requires a considerably larger number of samples (up to hundreds, while **WHAT** requires only few dozen).

3.4 Approach

3.4.1 Spectral Learning

The minimal sampling method proposed in this paper is based on a spectral-learning algorithm that explores the spectrum (eigenvalues) of the distance matrix between configurations. In theory, such spectral learners are an appropriate method to handle noisy, redundant, and tightly inter-connected variables, for the following reasons: When data sets have many irrelevancies or closely associated data parameters d, then only a few eigenvectors e, $e \ll d$ are required to characterize the data. In that reduced space:

- Multiple inter-connected variables $i, j, k \subseteq d$ can be represented by a single eigenvector;
- Noisy variables from d are ignored, because they do not contribute to the signal in the data;
- Variables become (approximately) parallel lines in e space. For redundancies $i, j \in d$, we can ignore j since effects that change over j also change in the same way over i;

That is, in theory, samples of configurations drawn via an eigenspace sampling method would not get confused by noisy, redundant, or tightly inter-connected variables. Accordingly, we expect predictions built from that sample to have lower mean errors and lower variances on that error.

Spectral methods have been used before for a variety of data mining applications [21]. Algorithms, such as PDDP [4], use spectral methods, such as principle component analysis (PCA), to recursively divide data into smaller regions. Software-analytics researchers use spectral methods (again, PCA) as a pre-processor prior to data mining to reduce noise in software-related data sets [42]. However, to the best of our knowledge, spectral methods have not been used before in software engineering as a basis of a minimal sampling method.

WHAT is somewhat different from other spectral learners explored in, for instance, image processing applications [36]. Work on image processing does not address defining a minimal sampling policy to predict performance scores. Also, a standard spectral method requires an $O(N^2)$ matrix multiplication to compute the components of PCA [18]. Worse, in the case of hierarchical division methods, such as PDDP, the polynomial-time inference must be repeated at every level of the hierarchy. Competitive results can be achieved using an $O(2N)$ analysis that we have developed previously [26], which is based on a heuristic proposed by Faloutsos and Lin [10] (which Platt has shown computes a Nyström approximation to the first component of PCA [31]).

Our approach inputs N (with $1 \leq |N| \leq |V|$) valid configurations ($\vec{C}$), $N_1, N_2, ...$, and then:

1. Picks any point N_i ($1 \leq i \leq |N|$) at random;
2. Finds the point $West \in N$ that is furthest away from N_i;
3. Finds the point $East \in N$ that is furthest from $West$.

The line joining *East* and *West* is our approximation for the first principal component. Using the distance calculation shown in Equation 3.2, we define δ to be the distance between *East* and *West*. **WHAT** uses this distance (δ) to divide all the configurations as follows: The value x_i is the projection of N_i on the line running from *East* to *West*[3]. We divide the examples based on

[3]The projection of N_i can be calculated in the following way:
$a = dist(East, N_i); b = dist(West, N_i); x_i = \sqrt{\frac{a^2 - b^2 + \delta^2}{2\delta}}$.

the median value of the projection of x_i. Now, we have two clusters of data divided based on the projection values (of N_i) on the line joining *East* and *West*. This process is applied recursively on these clusters until a predefined stopping condition. In our study, the recursive splitting of the N_i's stops when a sub-region contains less than $\sqrt{|N|}$ examples.

$$dist(x,y) = \begin{cases} \sqrt{\sum_i (x_i - y_i)^2} & \text{if } x_i \text{ and } y_i \text{ is numeric} \\ \begin{cases} 0, & \text{if } x_i = y_i \\ 1, & \text{otherwise} \end{cases} & \text{if } x_i \text{ and } y_i \text{ is Boolean} \end{cases} \tag{3.2}$$

We explore this approach for three reasons:

- *It is very fast*: This process requires only $2|n|$ distance comparisons per level of recursion, which is far less than the $O(N^2)$ required by PCA [8] or other algorithms such as K-Means [16].
- *It is not domain-specific*: Unlike traditional PCA, our approach is general in that it does not assume that all the variables are numeric. As shown in Equation 3.2,[4] we can approximate distances for both numeric and non-numeric data (e.g., Boolean).
- *It reduces the dimensionality problem*: This technique explores the underlying dimension (first principal component) without getting confused by noisy, related, and highly associated variables.

3.4.2 Spectral Sampling

When the above clustering method terminates, our sampling policy (which we will call S_1:Random) is then applied:

Random sampling (S_1): compile and execute one configuration, picked at random, from each leaf cluster;

We use this sampling policy, because (as we will show later) it performs better than:

East-West sampling (S_2): compile and execute the *East* and *West* poles of the leaf clusters;

Exemplar sampling (S_3): compile and execute all items in all leaves and return the one with lowest performance score.

[4] In our study, *dist* accepts configurations ($\vec{C}$) and returns the distance between them. If x_i and $y_i \in R^n$, then the distance function would be same as the standard Euclidean distance.

Note that S_3 is *not* a *minimal* sampling policy (since it executes all configurations). We use it here as one baseline against which we can compare the other, more minimal, sampling policies. In the results that follow, we also compare our sampling methods against another baseline using information gathered after executing all configurations.

3.4.3 Regression-Tree Learning

After collecting the data using one of the sampling policies (S_1, S_2, or S_3), as described in Section 3.4.2, we use a CART regression-tree learner [5] to build a performance predictor. Regression-tree learners seek the attribute-range split that most increases our ability to make accurate predictions. CART explores splits that divide N samples into two sets A and B, where each set has a standard deviation on the target variable of σ_1 and σ_2. CART finds the "best" split defined as the split that minimizes $\frac{A}{N}\sigma_1 + \frac{B}{N}\sigma_2$. Using this best split, CART divides the data recursively.

In summary, **WHAT** combines:

- The FASTMAP method of Faloutsos and Lin [10];
- A spectral-learning algorithm initially inspired by Boley's PDDP system [4], which we modify by replacing PCA with FASTMAP (called "WHERE" in prior work [26]);
- The sampling policy that explores the leaf clusters found by this recursive division;
- The CART regression-tree learner that converts the data from the samples collected by sampling policy into a run-time prediction model [5].

That is,

$$\begin{aligned} \text{WHERE} &= \text{PDDP} - \text{PCA} + \text{FASTMAP} \\ \textbf{WHAT} &= \text{WHERE} + \text{SamplingPolicy} + \text{CART} \end{aligned}$$

This unique combination of methods has not been previously explored in the software-engineering literature.

3.5 Experiments

All materials required for reproducing this work are available at https://goo.gl/689Dve.

3.5.1 Research Questions

We formulate our research questions in terms of the challenges of exploring large complex configuration spaces. Since our model explores the spectral space, our hypothesis is that only a small number of samples is required to explore the whole space. However, a prediction model built from a very small sample of the configuration space might be very inaccurate and unstable, that is, it may exhibit very large mean prediction errors and variances on the prediction error.

Also, if we learn models from small regions of the training data, it is possible that a learner will miss *trends* in the data between the sample points. Such trends are useful when building *optimizers* (i.e., systems that input one configuration and propose an alternate configuration that has, for instance, a better performance). Such optimizers might need to evaluate hundreds to millions of alternate configurations. To speed up that process, optimizers can use a *surrogate model*[5] that mimics the outputs of a system of interest, while being computationally cheap(er) to evaluate [24]. For example, when optimizing performance scores, we might ask a CART for a performance prediction (rather than compile and execute the corresponding configuration). Note that such surrogate-based reasoning critically depends on how well the surrogate can guide optimization.

Therefore, to assess feasibility of our sampling policies, we must consider:

- Performance scores generated from our minimal sampling policy;
- The variance of the error rates when comparing predicted performance scores with actual ones;
- The optimization support offered by the performance predictor (i.e., can the model work in tandem with other off-the-shelf optimizers to generate useful solutions).

The above considerations lead to four research questions:

RQ1: *Can* **WHAT** *generate good predictions after executing only a small number of configurations?*

Here, by "good" we mean that the predictions made by models that were trained using sampling with **WHAT** are as accurate, or more accurate, as those generated from models supplied with more samples.

RQ2: *Does less data used in building the models cause larger variances in the predicted values?*

RQ3: *Can "good" surrogate models (to be used in optimizers) be built from minimal samples?*

[5] Also known as response surface methods, meta models, or emulators.

Note that **RQ2** and **RQ3** are of particular concern with our approach, since our goal is to sample as little as possible from the configuration space.

RQ4: *How good is* **WHAT** *compared to the state of the art of learning performance predictors from configurable software systems?*

To answer RQ4, we will compare **WHAT** against approaches presented by Siegmund et al. [39], Guo et al. [15], and Sarkar et al. [33].

Berkeley DB C Edition (BDBC) is an embedded database system written in C. It is one of the most deployed databases in the world, due to its low binary footprint and its configuration abilities. We used the benchmark provided by the vendor to measure response time.
Berkeley DB Java Edition (BDBJ) is a complete re-development in Java with full SQL support. Again, we used a benchmark provided by the vendor measuring response time.
Apache is a prominent open-source Web server that comes with various configuration options. To measure performance, we used the tools autobench and httperf to generate load on the Web server. We increased the load until the server could not handle any further requests and marked the maximum load as the performance value.
SQLite is an embedded database system deployed over several millions of devices. It supports a vast number of configuration options in terms of compiler flags. As benchmark, we used the benchmark provided by the vendor and measured the response time.
LLVM is a compiler infrastructure written in C++. It provides various configuration options to tailor the compilation process. As benchmark, we measured the time to compile LLVM's test suite.
x264 is a video encoder in C that provides configuration options to adjust output quality of encoded video files. As benchmark, we encoded the Sintel trailer (735 MB) from AVI to the xH.264 codec and measured encoding time.

System	LOC	Features	Configurations
BDBC	219,811	18	2,560
BDBJ	42,596	32	400
Apache	230,277	9	192
SQLite	312,625	39	3,932,160
LLVM	47,549	11	1,024
x264	45,743	16	1,152

Figure 3.1 Subject systems used in the experiments.

3.5.2 Subject Systems

The configurable systems we used in our experiments are described in Figure 3.1. Note, with "predicting performance", we mean predicting performance scores of the subject systems while executing test suites provided by the developers or the community, as described in Figure 3.1. To compare the predictions of our and prior approaches with actual performance measures, we

use data sets that have been obtained by measuring *nearly all* configurations[6]. We say *nearly all* configurations, for the following reasoning: For all except one of our subject systems, the total number of valid configurations was tractable (192 to 2560). However, SQLite has 3,932,160 possible configurations, which is an impractically large number of configurations to test whether our predictions are accurate and stable. Hence, for SQLite, we use the 4500 samples for testing prediction accuracy and stability, which we could collect in one day of CPU time. Taking this into account, we will pay particular attention to the variance of the SQLite results.

3.5.3 Experimental Rig

RQ1 and **RQ2** require the construction and assessment of numerous runtime predictors from small samples of the data. The following rig implements that construction process.

For each configurable software system, we built a table of data, one row per valid configuration. We then ran all configurations of all software systems and recorded the performance scores (i.e., that are invoked by a benchmark). The exception is SQLite for which we measured only the configurations needed to detect interactions and additionally 100 random configurations to evaluate the accuracy of predictions. To this table, we added a column showing the performance score obtained from the actual measurements for each configuration.

Note that the following procedure ensures that we **never** test any prediction model on the data used to learn that model. Next, we repeated the following procedure 20 times (the figure of 20 repetitions was selected using the Central Limit Theorem): For each system in {BDBC, BDBJ, Apache, SQLite, LLVM, x264}

- Randomize the order of the rows in their table of data;
- For X in {10, 20, 30, ... , 90};
  - Let *Train* be the first X % of the data
  - Let *Test* be the rest of the data;
  - Pass *Train* to **WHAT** to select sample configurations;
  - Determine the performance scores associated with these configurations. This corresponds to a table lookup, but would entail compiling and executing a system configuration in a practical setting.
  - Using the *Train* data and their performance scores, build a performance predictor using CART.
  - Using the *Test* data, assess the accuracy of the predictor using the error measure of Equation 4.2 (see below).

[6] http://openscience.us/repo/performance-predict/cpm.html

The validity of the predictors built by the regression tree is verified on testing data. For each test item, we determine how long it *actually* takes to run the corresponding system configuration and compare the actual measured performance to the *prediction* from CART. The resulting prediction error is then computed using:

$$error = \frac{|\ predicted - actual\ |}{actual} * 100 \tag{3.3}$$

(Aside: It is reasonable to ask why this metrics and not some of the others proposed in the literature (e.g sum absolute residuals). In short, our results are stable across a range of different metrics. For example, the results of this paper have been repeated using sum of absolute residuals and, in those other results, we seen the same ranking of methods; see http://tiny.cc/sumAR.)

RQ2 requires testing the standard deviation of the prediction error rate. To support that test, we:

- Determine the X-th point in the above experiments, where all predictions stop improving (elbow point);
- Measure the standard deviation of the error at this point, across our 20 repeats.

As shown in Figure 3.2, all our results plateaued after studying $X = 40\,\%$ of the valid configurations[7]. Hence to answer **RQ2**, we will compare all 20 predictions at $X = 40\,\%$.

RQ3 uses the learned regression tree as a *surrogate model* within an optimizer;

- Take $X = 40\,\%$ of the configurations;
- Apply **WHAT** to build a CART model using some minimal sample taken from that $40\,\%$;
- Use that CART model within some standard optimizer while searching for configurations with least runtime;
- Compare the faster configurations found in this manner with the fastest configuration known for that system.

This last item requires access to a ground truth of performance scores for a large number of configurations. For this experiment, we have access to that ground truth (since we have access to all system configurations, except for SQLite). Note that such a ground truth would not be needed when practitioners choose to use **WHAT** in their own work (it is only for our empirical investigation).

[7]Just to clarify one frequently asked question about this work, we note that our rig "studies" $40\,\%$ of the data. We do not mean that our predictive models require accessing the performance scores from the $40\,\%$ of the data. Rather, by "study" we mean reflect on a sample of configurations to determine what minimal subset of that sample deserves to be compiled and executed.

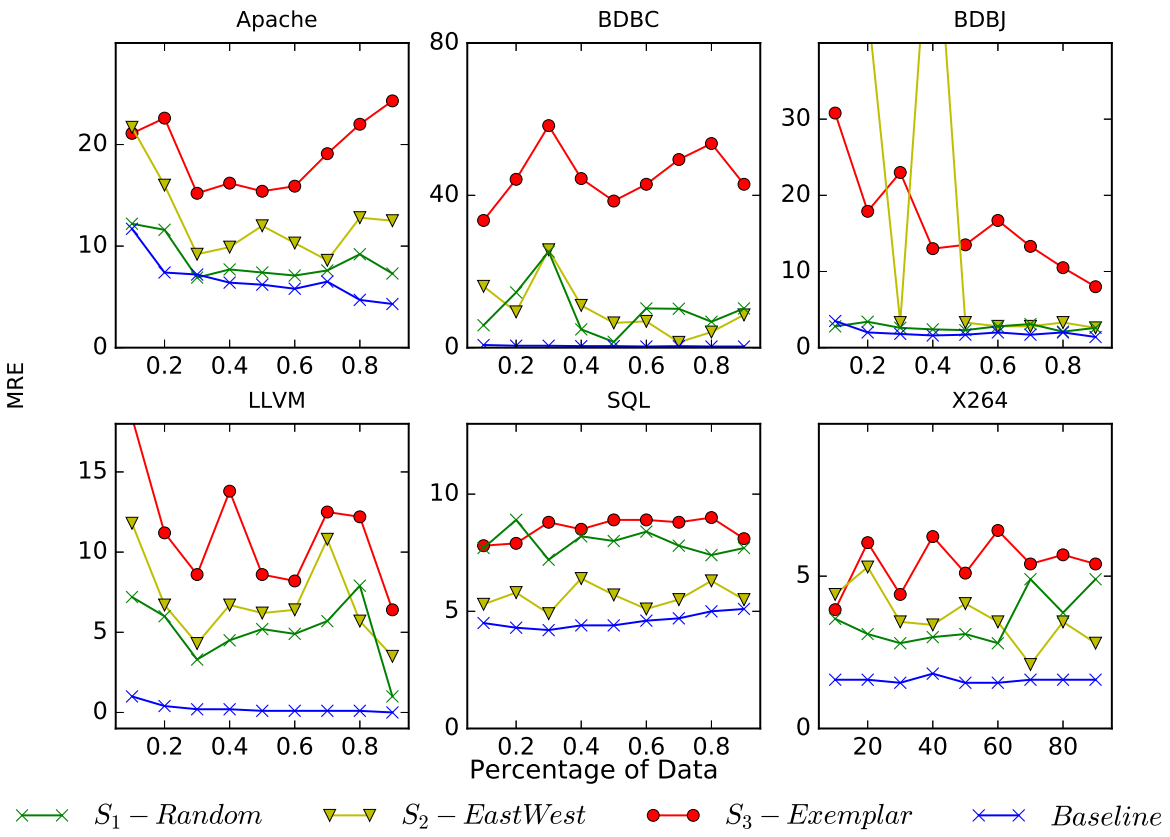

Figure 3.2 Errors of the predictions made by **WHAT** with four different sampling policies. Note that, on the y-axis, *lower* errors are *better*.

For the sake of completeness, we explored a range of optimizers seen in the literature in this second experiment: DE [41], NSGA-II [6], and our own GALE [22, 49] system. Normally, it would be reasonable to ask why we used those three, and not the hundreds of other optimizers described in the literature [12, 17]. However, as shown below, all these optimizers in this domain exhibited very similar behavior (all found configurations close to the best case performance). Hence, the specific choice of optimizer is not a critical variable in our analysis.

3.6 Results

3.6.1 RQ1

Can **WHAT** *generate good predictions after executing only a small number of configurations?*

Figure **??** shows the mean errors of the predictors learned after taking X % of the configurations, then asking **WHAT** and some sampling method (S_1, S_2, and S_3) to (a) find what configurations

to measure; then (b) asking CART to build a predictor using these measurements. The horizontal axis of the plots shows what X % of the configurations are studied; the vertical axis shows the mean relative error (from Equation 4.2). In that figure:

- The ×—× lines in Figure **??** show a *baseline* result where data from the performance scores of 100 % of configurations were used by CART to build a runtime predictor.
- The other lines show the results using the sampling methods defined in Section 3.4.2. Note that these sampling methods used runtime data only from a subset of 100 % of the performance scores seen in configurations from 0 to X %.

In Figure 3.2, *lower* y-axis values are *better* since this means lower prediction errors. Overall, we find that:

- Some software systems exhibit large variances in their error rate, below $X = 40$ % (e.g., BDBC and BDBJ).
- Above $X = 40$ %, there is little effect on the overall change of the sampling methods.
- Mostly, S_3 shows the highest overall error, so that it cannot be recommended.
- Always, the ×—× baseline shows the lowest errors, which is to be expected since predictors built on the baseline have access to all data.
- We see a trend that the error of S_1 and S_2 are within 5 % of the *baseline* results. Hence, we can recommend these two minimal sampling methods.

Figure 3.3 provides information about which of S_1 or S_2 we should recommend. This figure displays data taken from the $X = 40$ % point of Figure 3.2 and displays how many performance scores of configurations are needed by our sub-sampling methods (while reflecting on the configurations seen in the range $0 \leq X \leq 40$). Note that:

- S_3 needs up to thousands of performance-score points, so it cannot be recommended as minimal-sampling policy;
- S_2 needs twice as much performance-score information as S_1 (S_2 uses *two* samples per leaf cluster while S_1 uses only *one*).
- S_1 needs performance-score information on only a few dozen (or less) configurations to generate the predictions with the lower errors seen in Figure 3.2.

Combining the results of Figure 3.2 and Figure 3.3, we conclude that:

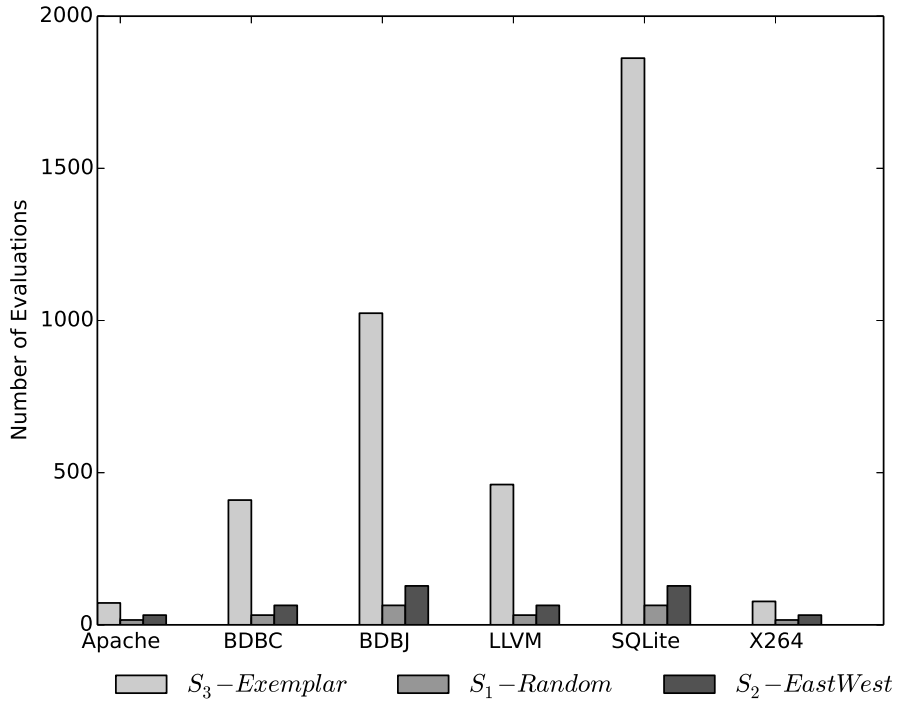

Figure 3.3 Comparing evaluations of different sampling policies. We see that the number of configurations evaluated for S_2 is twice as high as S_1, as it selects 2 points from each cluster, where as S_1 selects only 1 point.

S_1 is our preferred spectral sampling method. Furthermore, the answer to **RQ1** is "yes", because applying **WHAT** , we can (a) generate runtime predictors using just a few dozens of sample performance scores; and (b) these predictions have error rates within 5 % of the error rates seen if predictors are built from information about all performance scores.

3.6.2 RQ2

Do less data used in building models cause larger variances in the predicted values?

Two competing effects can cause increased or decreased variances in runtime predictions. The less we sample the configuration space, the less we constrain model generation in that space. Hence, one effect that can be expected is that models learned from too few samples exhibit large variances. But, a compensating effect can be introduced by sampling from the spectral space since that space contains fewer confusing or correlated variables than the raw configuration space. Figure 3.4 reports which one of these two competing effects are dominant.

Figure 3.2 shows that after some initial fluctuations, after seeing $X = 40\,\%$ of the configurations, the variances in prediction errors reduces to nearly zero.

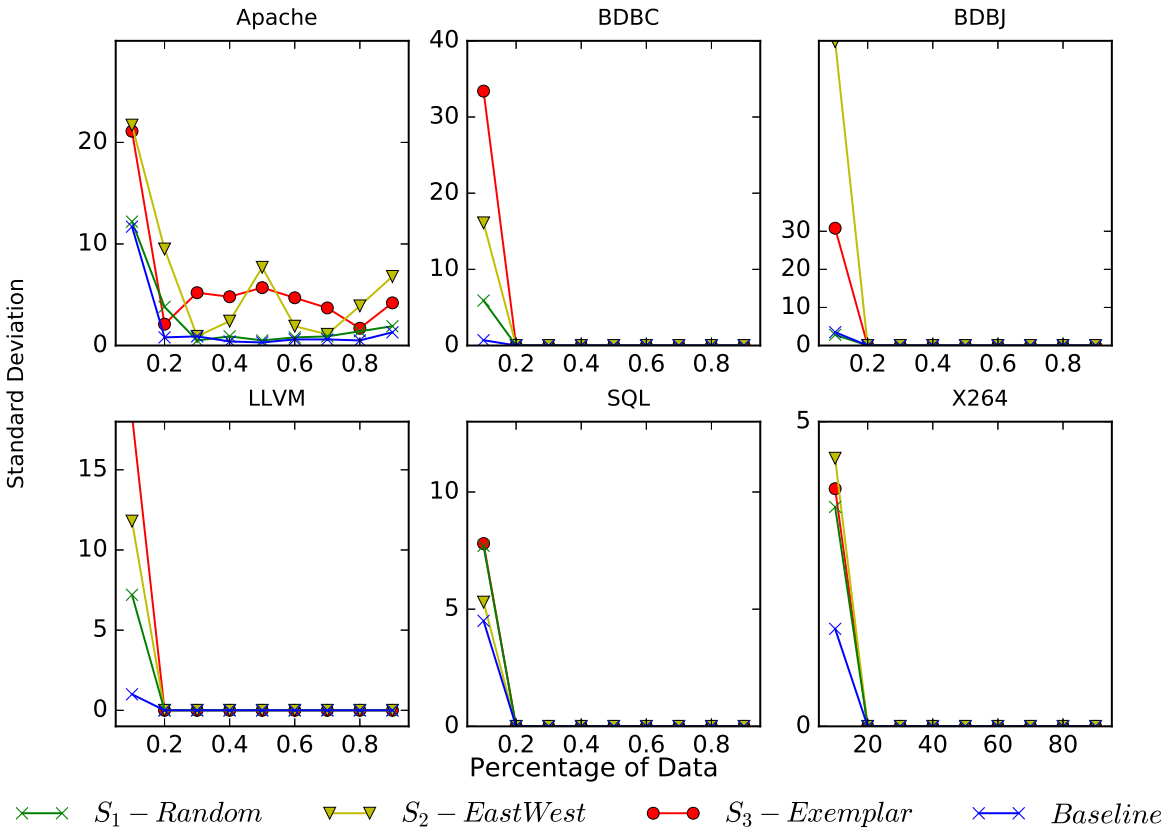

Figure 3.4 Standard deviations seen at various points of Figure 3.2.

Hence, we answer **RQ2** with "no": Selecting a small number of samples does not necessarily increase variance (at least to say, not in this domain).

3.6.3 RQ3

Can "good" surrogate models (to be used in optimizers) be built from minimal samples?

The results of answering **RQ1** and **RQ2** suggest to use **WHAT** (with S_1) to build runtime predictors from a small sample of data. **RQ3** asks if that predictor can be used by an optimizer to infer what *other* configurations correspond to system configurations with fast performance scores. To answer this question, we ran a random set of 100 configurations, 20 times, and related

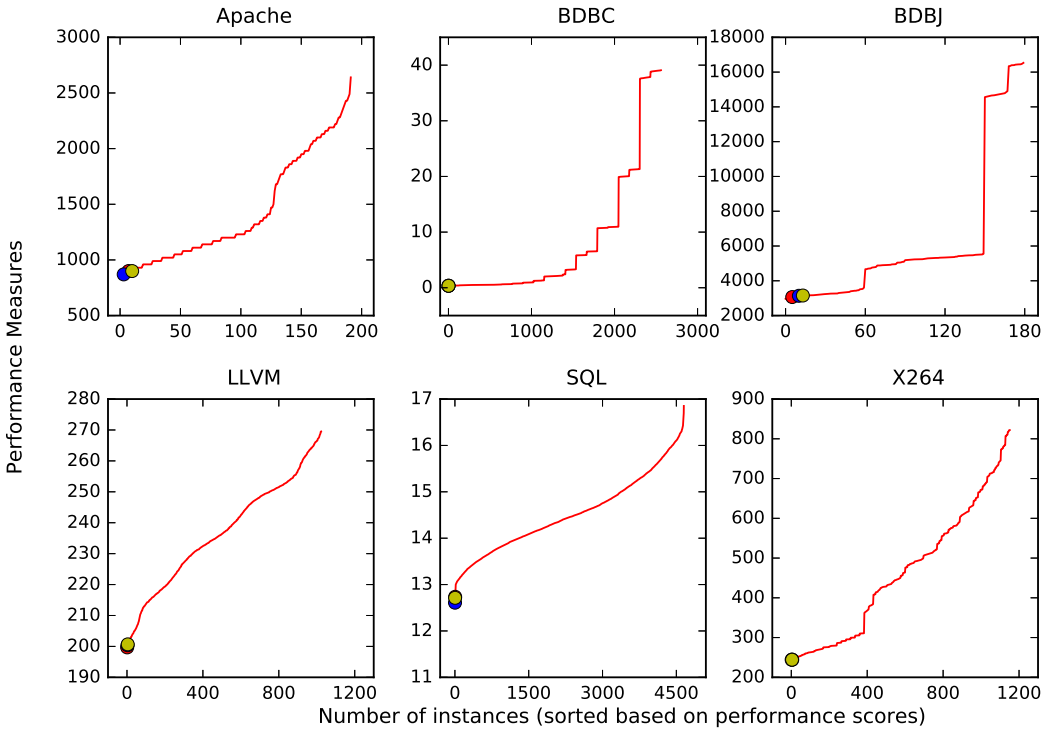

Figure 3.5 Solutions found by GALE, NSGA-II and DE (shown as points) laid against the ground truth (all known configuration performance scores). It can be observed that all the optimizers can find the configuration with lower performance scores.

that baseline to three optimizers (GALE [22], DE [41] and NSGA-II [6]) using their default parameters.

When these three optimizers mutated existing configurations to suggest new ones, these mutations were checked for validity. Any mutants that violated the system's constraints (e.g., a feature excluding another feature) were rejected and the survivors were "evaluated" by asking the CART surrogate model. These evaluations either rejected the mutant or used it in generation $i + 1$, as the basis for a search for more, possibly better mutants.

Figure 3.5 shows the configurations found by three optimizers projected onto the ground truth of the performance scores of nearly all configurations (see Section 3.5.2). Again note that, while we use that ground truth for the validation of these results, our optimizers used only a small part of that ground-truth data in their search for the fastest configurations (see the **WHAT** + S_1 results of Figure 3.3).

The important feature of Figure 3.5 is that all the optimized configurations fall within 1 % of the fastest configuration according to the ground truth (see all the left-hand-side dots on each

plot). Table 3.1 compares the performance of the optimizers used in this study. Note that the performances are nearly identical, which leads to the following conclusions:

The answer to **RQ3** is "yes": For optimizing performance scores, we can use surrogates built from few runtime samples. The choice of the optimizer does not critically effect this conclusion.

Table 3.1 The table shows how the minimum performance scores as found by the learners GALE, NSGA-II, and DE, vary over 20 repeated runs. Mean values are denoted μ and IQR denotes the 25th–75th percentile. A low IQR suggests that the surrogate model build by **WHAT** is stable and can be utilized by off the shelf optimizers to find performance-optimal configurations.

Dataset	**Searcher**					
	GALE		**DE**		**NSGAII**	
	Mean	**IQR**	**Mean**	**IQR**	**Mean**	**IQR**
Apache	870	0	840	0	840	0
BDBC	0.363	0.004	0.359	0.002	0.354	0.005
BDBJ	3139	70	3139	70	3139	70
LLVM	202	3.98	200	0	200	0
SQLite	13.1	0.241	13.1	0	13.1	0.406
X264	248	3.3	244	0.003	244	0.05

3.6.4 RQ4

How good is **WHAT** *compared to the state of the art of learning performance predictors from configurable software systems?*

We compare **WHAT** with the three state-of-the-art predictors proposed in the literature [39], [15], [33], as discussed in Section 3.3. Note that all approaches use regression-trees as predictors, except Siegmund's approach, which uses a regression function derived using linear programming. The results were studied using non-parametric tests, which was also used by Arcuri and Briand at ICSE '11 [27]). For testing statistical significance, we used non-parametric bootstrap test 95% confidence [9] followed by an A12 test to check that any observed differences were not trivially small effects; i.e. given two lists X and Y, count how often there are larger numbers in the former list (and there there are ties, add a half mark): $a = \forall x \in X, y \in Y \frac{\#(x>y)+0.5*\#(x=y)}{|X|*|Y|}$ (as

per Vargha [43], we say that a "small" effect has $a < 0.6$). Lastly, to generate succinct reports, we use the Scott-Knott test to recursively divide our optimizers. This recursion used A12 and bootstrapping to group together subsets that are (a) not significantly different and are (b) not just a small effect different to each other. This use of Scott-Knott is endorsed by Mittas and Angelis [27] and by Hassan et al. [13].

As seen in the Figure 3.6, the FW approach of Siegmund et al. (i.e., the sampling approach using the fewest number of configurations) (4/6) times has the higher errors rate and the highest standard deviation on that error rate. Hence, we cannot recommend this method or, if one wishes to use this method, we recommend using the other sampling heuristics (e.g., HO, HS) to make more accurate predictions (but at the cost of much more measurements). Moreover, the size of the standard deviation of this method causes further difficulties in estimating which configurations are those exhibiting a large prediction error.

THere are two cases in Figure 3.6 where **WHAT** performs worse than at least one other methods:

- SQLite: here, the Sukar methods does better that **WHAT** (3.44 vs 5.6) but, as shown in the final column of Figure 3.6, does so at the cost of $\frac{925}{64} \approx 15$ times more evaluations that **WHAT** . In this case, a pragmatic engineering could well prefer our solution the that of Sukar (since Sukar uses more than an order of magnitude slowdown).

- BDBC: Here again, **WHAT** is not doing the best but, compared to the space of all other solutions, it is not doing particular bad.

As to the approach of Guo et al (with PW), this does not standout on any of our measurements. Its error results are within 1% of **WHAT** ; its standard deviation are usually larger and it requires much more data than **WHAT** (Evaluations column of the figure).

In terms of the number of measure samples required to build a model, the right-hand most column of Figure 3.6 shows that **WHAT** requires the fewest samples except for two cases: the approach of Guo et al. (with 2N) working on BDBC and LLVM. In both these cases, the mean error and standard deviation on the error estimate is larger than **WHAT** . Furthermore, in the case of BDBC, the error values are $\mu = 14\,\%$, $\sigma = 13\,\%$, which are much larger than **WHAT** 's error scores of $\mu = 6\,\%$, $\sigma = 5\,\%$.

Although the approach of Sarkar et al. produces an error rate that is sometimes less than the one of **WHAT** , it requires the most number of measurements. Moreover, **WHAT** ś accuracy is close to Sarkarś approach (1% to 2%) difference). Hence, we cannot recommend this approach, too.

Table 3.2 shows the number of evaluations used by each approaches. We see that most state-of-the-art approaches often require many more samples than **WHAT** . Using those fewest

Table 3.2 Comparison of the number of the samples required with the state of the art. The grey colored cells indicate the approach which has the lowest number of samples. We notice that WHAT and Guo (2N) uses less data compared to other approaches. The high fault rate of Guo (2N) accompanied with high variability in the predictions makes WHAT our preferred method.

Dataset	Samples				
	Siegmund	**Guo (2N)**	**Guo (PW)**	**Sarkar**	**WHAT**
Apache	29	181	29	55	16
BDBC	139	36	139	191	64
BDBJ	48	52	48	57	16
LLVM	62	22	64	43	32
SQLite	566	78	566	925	64
X264	81	32	81	93	32

numbers of samples, **WHAT** has within 1 to 2 % of the lowest standard deviation rates and within 1 to 2 % of lowest error rates. The exception is Sarkar's approach, which has 5 % lower mean error rates (in BDBC, see the left-hand-side plot of Figure **??**). However, as shown in right-hand-side of Table 3.2, Sarkar's approach needs nearly three times more measurements than **WHAT** (191 vs 64 samples). Given the overall reduction of the error is small (5 % difference between Sarkar and **WHAT** in mean error), the overall cost of tripling the data-collection cost is often not feasible in a practical context and might not justify the small additional benefit in accuracy.

Hence, we answer **RQ4** with "yes", since **WHAT** yields predictions that are similar to or more accurate than prior work, while requiring fewer samples.

3.7 Why does it work?

In this section, we present an in-depth analysis to understand why our sampling technique (based on a spectral learner) achieves such low mean fault rates while being stable (low variance). We hypothesize that the configuration space of the system configuration lie on a low dimensional manifold.

3.7.1 History

Menzies et. al [26] suggested data heterogeneity by demonstrating the existence of local regions, which were very different from the global representation. Menzies et al. also makes an important

observation about exploiting the underlying dimension to cluster the data. The authors used an algorithm called WHERE (see section 3.4.3), which recurses on two dimensions synthesized in linear time using a technique called FASTMAP [11]. The use of underlying dimension has been endorsed by various other researchers [1, 2, 7, 47]. There are numerous other methods in the literature, which are used to learn the underlying dimensionality of the data set like Principal Component Analysis (PCA) [20] [8], Spectral Learning [37], Random Projection [3]. These algorithms use different techniques to identify the underlying, independent/orthogonal dimensions to cluster the data points and differ with respect to the computational complexity and accuracy. We use WHERE since it computationally efficient $O(2N)$ while being accurate.

3.7.2 Testing Technique

Given our hypothesis – configuration space lies in a lower dimensional hyperplane– it is imperative to demonstrate that the intrinsic dimensionality of the configuration space is less than actual dimension. To formalize this notion we borrow the concept of correlation dimension from the domain of physics [14]. The correlation dimension of a dataset with k item is found by computing the number of items found at distance withing radius r (where r is the euclidean distance between two configurations) while varying r. This is then normalized by the number of connections between k items to find the expected number of neighbors at distance r. This can be written as:

$$C(r) = \frac{2}{k(k-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} I(||x_i, x_j|| < r) \tag{3.4}$$

$$where : I(x < y) = \begin{cases} 1, & \text{if x <y} \\ 0, & \text{otherwise} \end{cases}$$

Given the dataset with k items and range of distances [r_0–r_{max}], we estimate the intrinsic dimensionality as the maximum slope between $\ln(C(r))$ vs $\ln(r)$.

3.7.3 Evaluation

We observe that **the intrinsic dimensionality of the software system is much lower than the actual dimension**. Figure 3.7 presents the intrinsic dimensionality along with the actual dimensions of the software systems. If we take a look at the intrinsic dimensionality and compare it with the actual dimensionality, then it becomes apparent that the configuration space lies on a lower dimensional hyperplane. For example, SQLite has 39 configuration options but the intrinsic dimensionality of the space is just 7.61 (this is a fractal dimension). At the

[8] WHERE is an approximation of the first principal component

heart of **WHAT** is WHERE (a spectral clusterer), which uses the approximation of the first principal component to divide the configuration space and hence can take advantage of the low intrinsic dimensionality.

As a summary, our observations indicates that the intrinsic dimension of the configuration space is much lower that its actual dimension. Hence, clustering based on the intrinsic dimensions rather than the actual dimension would be more effective. In other words, configurations with similar performance values lie closer in the intrinsic hyperplane when compared to the actual dimensions and may be the reason as to why **WHAT** achieves empirically good results.

> The intrinsic dimension of the configuration space is much lower than its actual dimension and hence clustering based on the intrinsic dimension is more effective.

3.8 Reliability and Validity

Reliability refers to the consistency of the results obtained from the research. For example, how well independent researchers could reproduce the study? To increase external reliability, this paper has taken care to either clearly define our algorithms or use implementations from the public domain (SciKitLearn) [35]. Also, all the data used in this work is available on-line in the PROMISE code repository and all our algorithms are on-line at github.com/ai-se/where.

Validity refers to the extent to which a piece of research actually investigates what the researcher purports to investigate [38]. *Internal validity* checks if the differences found in the treatments can be ascribed to the treatments under study.

One internal validity issue with our experiments is the choice of *training and testing* data sets discussed in Figure 3.1. Recall that while all our learners used the same *testing* data set, our untuned learners were only given access to *training* data.

Another internal validity issues is *instrumentation*. The very low μ and σ error values reported in this study are so small that it is reasonable to ask whether they are due to some instrumentation quirk, rather than due to using a clever sample strategy:

- Our low μ values are consistent with prior work (e.g. [33]);
- As to our low σ values, we note that, when the error values are so close to 0 %, the standard deviation of the error is "squeezed" between zero and those errors. Hence, we would expect that experimental rigs that generate error values on the order of 5 % and Equation 4.2 should have σ values of $0 \leq \sigma \leq 5$ (e.g., like those seen in our introduction).

Regarding SQLite, we cannot measure all possible configurations in reasonable time. Hence, we sampled only 100 configurations to compare prediction and actual performance values. We

are aware that this evaluation leaves room for outliers. Also, we are aware that measurement bias can cause false interpretations [26]. Since we aim at predicting performance for a special workload, we do not have to vary benchmarks.

We aimed at increasing the *external validity* by choosing software systems from different domains with different configuration mechanisms and implemented with different programming languages. Furthermore, the systems used are deployed and used in the real world. Nevertheless, assuming the evaluations to be automatically transferable to all configurable software systems is not fair. To further strengthen external validity, we run the model (generated by **WHAT** + S_1) against other optimizers, such as NSGA-II and differential evolution [41]. That is, we validated whether the learned models are not only applicable for GALE style of perturbation. In Table 3.1, we see that the models developed are valid for all optimizers, as all optimizers are able to find the near optimal solutions.

3.9 Related Work

In 2000, Shi and Maik [36] claimed the term "spectral clustering" as a reference to their normalized cuts image segmentation algorithm that partitions data through a spectral (eigenvalue) analysis of the Laplacian representation of the similarity graph between instances in the data.

In 2003, Kamvar et al. [21] generalized that definition saying that "spectral learners" were any data-mining algorithm that first replaced the raw dimensions with those inferred from the spectrum (eigenvalues) of the affinity (a.k.a. distance) matrix of the data, optionally adjusted via some normalization technique).

Our clustering based on first principal component splits the data on a approximation to an eigenvector, found at each recursive level of the data (as described in §3.4.1). Hence, this method is a "spectral clusterer" in the general Kamvar sense. Note that, for our data, we have not found that Kamvar's normalization matrices are needed.

Regarding sampling, there are a wide range of methods know as experimental designs or designs of experiments [32]. They usually rely on fractional factorial designs as in the combinatorial testing community [23].

Furthermore, there is a recent approach that learns *performance-influence models* for configurable software systems [40]. While this approach can handle even numeric features, it has similar sampling techniques for the Boolean features as reported in their earlier work [39]. Since we already compared to that earlier work and do not consider also numeric features, we did not compare our work to performance-influence models.

3.10 Conclusions

Configurable software systems today are widely used in practice, but expose challenges regarding finding performance-optimal configurations. State-of-the-art approaches require too many measurements or are prone to large variances in their performance predictions. To avoid these shortcomings, we have proposed a fast spectral learner, called **WHAT** , along with three new sampling techniques. The key idea of **WHAT** is to explore the configuration space with eigenvalues of the features used in a configuration to determine exactly those configurations for measurement that reveal key performance characteristics. This way, we can study many closely associated configurations with only a few measurements.

We evaluated our approach on six real-world configurable software systems borrowed from the literature. Our approach achieves similar to lower error rates, while being stable when compared to the state of the art. In particular, with the exception of Berkeley DB, our approach is more accurate than the state-of-the-art approaches by Siegmund et al. [39] and Guo et al. [15]. Furthermore, we achieve a similar prediction accuracy and stability as the approach by Sarkar et al [33], while requiring a far smaller number of configurations to be measured. We also demonstrated that our approach can be used to build cheap and stable surrogate prediction models, which can be used by off-the-shelf optimizers to find the performance-optimal configuration.

Rank	using	Mean MRE	STDev	Evaluations
Apache				
1	Sarkar	7.49	0.82	55
1	Guo(PW)	10.51	6.85	29
1	Siegmund	10.34	11.68	29
1	**WHAT**	10.95	2.74	16
1	Guo(2N)	13.03	15.28	18
BDBC				
1	Sarkar	1.24	1.46	191
2	Siegmund	6.14	4.41	139
2	**WHAT**	6.57	7.4	64
2	Guo(PW)	10.16	10.6	139
3	Guo(2N)	49.9	52.25	36
BDBJ				
1	Guo(2N)	2.29	3.26	52
1	Guo(PW)	2.86	2.72	48
1	**WHAT**	4.75	4.46	16
2	Sarkar	5.67	6.97	48
2	Siegmund	6.98	7.13	57
LLVM				
1	Guo(PW)	3.09	2.98	64
1	**WHAT**	3.32	1.05	32
1	Sarkar	3.72	0.45	62
1	Guo(2N)	4.99	5.05	22
2	Siegmund	8.5	8.28	43
SQLite				
1	Sarkar	3.44	0.1	925
2	**WHAT**	5.6	0.57	64
3	Guo(2N)	8.57	7.3	78
3	Guo(PW)	8.94	6.24	566
4	Siegmund	12.83	17.0	566
x264				
1	Sarkar	6.64	1.04	93
1	**WHAT**	6.93	1.67	32
1	Guo(2N)	7.18	7.07	32
1	Guo(PW)	7.72	2.33	81
2	Siegmund	31.87	21.24	81

Figure 3.6 Mean MRE seen in 20 repeats. Mean MRE is the prediction error as described in Equation 4.2 and STDev is the standard deviation of the MREs found during multiple repeats. Lines with a a dot in the middle (e.g. —•) show the mean as a round dot withing the IQR (and if the IQR is very small, only a round dot will be visible). All the results are sorted by the mean values: lower mean value of MRE is better than large mean value. The left-hand side columns **Rank** the various techniques, smaller the value of **Rank** better the technique e.g. in Apache, all the techniques have the same rank since their mean values are not statistically different. **Rank** is computer using Scott-Knott, bootstrap 95% confidence, and the A12 test.

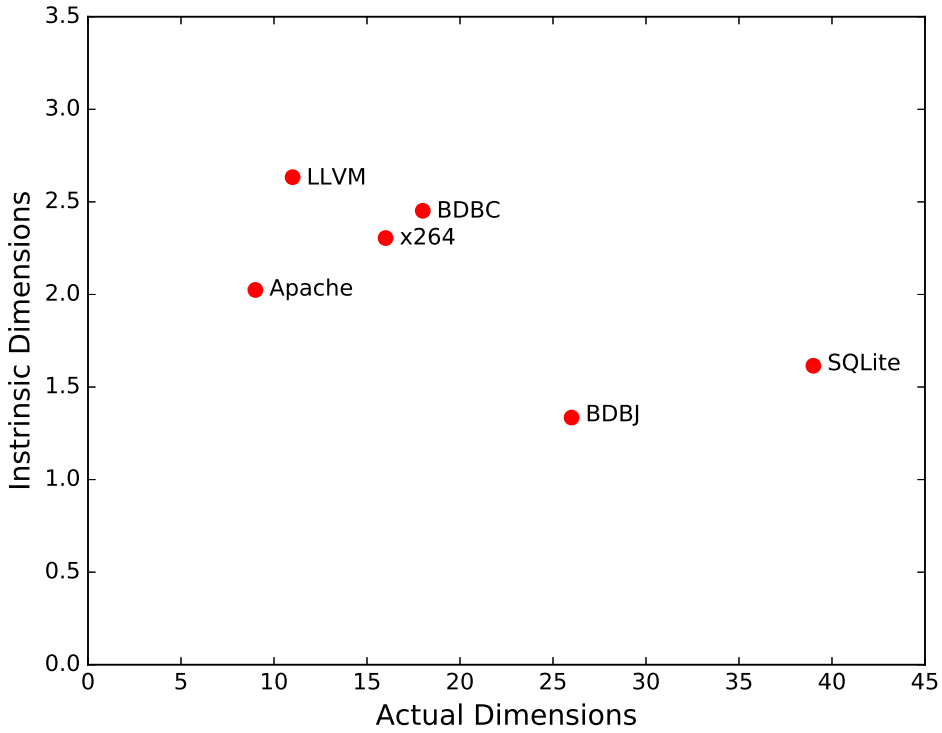

Figure 3.7 Intrinsic dimensionality of the subjects systems are shown on the y-axis. The number on the side is the actual dimension of the system. The intrinsic dimensionality of the systems are much lower than the actual dimensionality (number of columns in the dataset).

Chapter 4

Using Bad Learners to find Good Configurations

This chapter originally appeared as Vivek Nair, Tim Menzies, Norbert Siegmund, and Sven Apel. 2017. Using bad learners to find good configurations. In Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2017).

4.1 Abstract

Finding the optimally performing configuration of a software system for a given setting is often challenging. Recent approaches address this challenge by learning performance models based on a sample set of configurations. However, building an accurate performance model can be very expensive (and is often infeasible in practice). The central insight of this paper is that exact performance values (e.g., the response time of a software system) are not required to rank configurations and to identify the optimal one. As shown by our experiments, performance models that are cheap to learn but inaccurate (with respect to the difference between actual and predicted performance) can still be used rank configurations and hence find the optimal configuration. This novel *rank-based approach* allows us to significantly reduce the cost (in terms of number of measurements of sample configuration) as well as the time required to build performance models. We evaluate our approach with 21 scenarios based on 9 software systems and demonstrate that our approach is beneficial in 16 scenarios; for the remaining 5 scenarios, an accurate model can be built by using very few samples anyway, without the need for a rank-based approach.

4.2 Introduction

This paper proposes an improvement of recent papers presented at ICSE'12, ASE'13, and ASE'15, which predict system performance based on learning influences of individual configuration options and combinations of thereof [15, 33, 39]. The idea is to measure a few configurations of a configurable software system and to make statements about the performance of its other configurations. Thereby, the goal is to predict the performance of a given configuration as accurate as possible. We show that, if we (slightly) relax the question we ask, we can build useful predictors using very small sample sets. Specifically, instead of asking "How long will this configuration run?", we ask instead "Will this configuration run faster than that configuration?" or "Which is the fastest configuration?".

This is an important area of research since understanding system configurations has become a major problem in modern software systems. In their recent paper, Xu et al. documented the difficulties developers face with understanding the configuration spaces of their systems [46]. As a result, developers tend to ignore over 83% of configuration options, which leaves considerable optimization potential untapped and induces major economic cost [46].

With many configurations available for today's software systems, it is challenging to optimize for functional and non-functional properties. For functional properties, Chen et. al [**chen2016sampling**] and Sayyad et. al [**sayyad2013scalable**] developed fast techniques to find near-optimal configurations by solving a five-goal optimization problem. Henard et. al [**henard2015combining**] used a SAT solver along with Multi-Objective Evolutionary Algorithms to repair invalid mutants found during the search process.

For non-functional properties, researchers have also developed a number of approaches. For example, it has been shown that the runtime of a configuration can be predicted with high accuracy by sampling and learning performance models [15, 33, 39]. State-of-the-art techniques rely on configuration data from which it is possible to build very accurate models. For example, prior work [28] has used sub-sampling to build predictors for configuration runtimes using predictors with error rates less than 5% (quantified in terms of *residual-based* measures such as Mean Magnitude of Relative Error, or MMRE, $(\sum_i^n(|a_i - p_i|/a_i))/n$ where a_i, p_i are the *actual* and *predicted values*). Figure 4.1 shows in green a number of real-world systems whose performance behavior can be modelled with high accuracy using state-of-the-art techniques.

Recently, we have come across software systems whose configuration spaces are far more complicated and hard to model. For example, when the state-of-the-art technique of Guo at al. [15] is applied to these software systems, the error rates of the generated predictor is up to 80%—see the yellow and **red** systems of Figure 4.1. The existence of these harder-to-model systems raises a serious validity question for all prior research in this area:

- Was prior research merely solving easy problems?

- Can we learn predictors for non-functional properties of more complex systems?

One pragmatic issue that complicates answering these two questions is the *minimal sampling problem*. It can be prohibitively expensive to run and test all configurations of modern software systems since their configuration spaces are very large. For example, to obtain the data used in our experiments, we required over a month of CPU time for measuring (and much longer, if we also count the time required for compiling the code prior to execution). Other researchers have commented that, in real-world scenarios, the cost of acquiring the optimal configuration is overly expensive and time-consuming [45]. Hence, the goal of this paper must be:

1. Find predictors for non-functional properties for the hard-to-model systems of Figure 4.1, where learning accurate performance models is expensive.

2. Use as few sample configurations as possible.

The key result of this paper is that, even when *residual-based* performance models are inaccurate, *ranking* performance models can still be very useful for configuration optimization. Note that:

- Predictive models return a value for a configuration;

- Ranking models rank N configurations from "best" to "worst".

There are two very practical cases where such ranking models would suffice:

- Developers want to know the fastest configuration;

- Developers are debating alternate configurations and want to know which might run faster.

In this paper, we explore two research questions about constructing ranking models.
RQ1: *Can inaccurate predictive models still accurately rank configurations?*

We show below that, even if a model has, overall, a low predictive accuracy (i.e., a high MMRE), the predictions can still be used to effectively rank configurations. The rankings are heuristic in nature and hence may be slightly inaccurate (w.r.t the actual performance value). That said, overall, our rankings are surprisingly accurate. For example, when exploring the configuration space of SQLite, our rankings are usually wrong only by less than 6 neighboring configurations – which is a very small number considering that SQLite has almost 4 million configurations.
RQ2: *How expensive is a rank-based approach (in terms of how many configurations must be executed)?*

To answer this question, we studied the configurations of 21 scenarios based on 9 open-source systems. We measure the benefit of our rank-based approach as the percentage of required

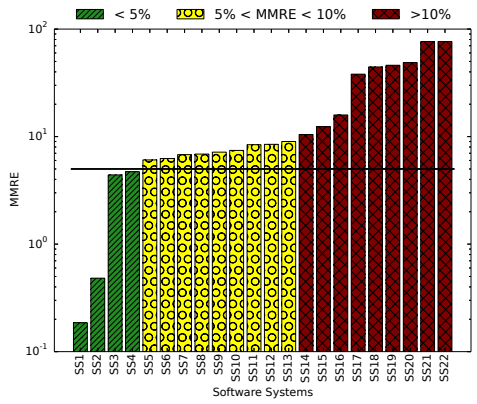

Figure 4.1 Errors of the predictions made by using CART, a machine learning technique (refer to Section 4.6), to model different software systems. Due to the results of Figure 4.2, we use 30%/70% of the valid configurations (chosen randomly) to train/test the model.

measurements needed by state-of-the-art techniques in this field (see Sarkar et al. [33] presented at ASE'15). Those percentages were as follows – Note that *lower* values are *better* and values under 100% denote an improvement over the state-of-the-art (from Figure 4.9):

$$\{5,5,5,5,5,10,20,20,20,20,30,30,35,35,40,40,50,50,70,80,80,110\}\%$$

That is, the novel rank-based approach described in this paper is rarely worse than the state of the art and often far better. For example, as shown later in Figure 4.9, for one of the scenarios of Apache Storm, SS11, the rank-based approach uses only 5% of the measurements used by a residual-based approach.

The rest of this paper is structured as follows: we first formally describe the prediction problem. Then, we describe the state-of-the-art approach proposed by Sarkar et al. [33], henceforth referred to as residual-based approach, followed by the description of our rank-based approach. Then, the subject systems used in the paper are described followed by our evaluation. The paper ends with a discussion on why a rank-based approach works; finally, we conclude. To assist other researchers, a reproduction package with all our scripts and data are available on GitHub.[1]

4.3 Problem Formalization

A configurable software system has a set X of configurations $x \in X$. Let x_i indicate the *ith* configuration option of configuration x, which takes the values from a finite domain $Dom(x_i)$. In general, x_i indicates either an (i) integer variable or a (ii) Boolean variable. The configuration

[1] https://github.com/ai-se/Reimplement/tree/cleaned_version

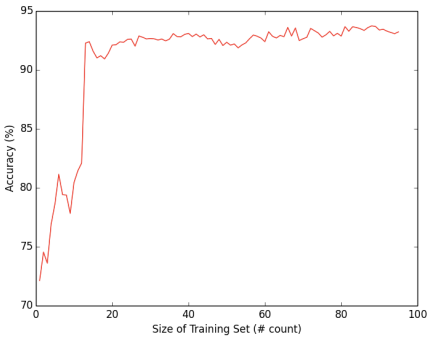

Figure 4.2 The relationship between the accuracy (in terms of MMRE) and the number of samples used to train the performance model of the Apache Web Server. Note that the accuracy does not improve substantially after 20 sample configurations.

space is thus $Dom(x_1) \times Dom(x_2) \times ... \times Dom(x_n)$, which is the Cartesian product of the domains, where $n = |x|$ is the number of configuration options of the system. Each configuration (x) has a corresponding performance measure $y \in Y$ associated with it. The performance measure is also referred to as dependent variable. We denote the performance measure associated with a given configuration by $y = f(x)$. We consider the problem of ranking configurations (x^*) that such that $f(x)$ is less than other configurations in the configuration space of X with few measurements.

$$f(x^*) \leq f(x), \quad \forall x \in X \setminus x^* \tag{4.1}$$

Our goal is to find the (near) optimal configuration of a system where it is not possible to build an accurate performance model as prescribed in earlier work.

4.4 Residual-based Approaches

In this section, we discuss the residual-based approaches for building performance models for configurable software systems. For further details, we refer to Sarkar et. al [33].

4.4.1 Progressive Sampling

When the cost of collecting data is higher than the cost of building a performance model, it is imperative to minimize the number of measurements required for model building. A learning curve shows the relationship between the size of the training set and the accuracy of the model. In Figure 4.2, the horizontal axis represents the number of samples used to create the performance model, whereas the vertical axis represents the accuracy (measured in terms of MMRE) of the model learned. Learning curves typically have a steep sloping portion early in the curve followed

```
1    # Progressive Sampling
2    def progressive(training, testing, lives=3):
3      # For stopping criterion
4      last_score = -1
5      independent_vals = list()
6      dependent_vals = list()
7      for count in range(1, len(training)):
8        # Add one configuration to the training set
9        independent_vals += training[count]
10       # Measure the performance value for the newly
11       # added configuration
12       dependent_vals += measure(training_set[count])
13       # Build model
14       model = build_model(independent_vals, dependent_vals)
15       # Test Model
16       perf_score = test_model(model, testing, measure(testing))
17       # If current accuracy score is not better than
18       # the previous accuracy score, then loose life
19       if perf_score <= last_score:
20         lives -= 1
21       last_score = perf_score
22       # If all lives are lost, exit loop
23       if lives == 0: break
24     return model
```

Figure 4.3 Pseudocode of progressive sampling.

by a plateau late in the curve. The plateau occurs when adding data does not improve the accuracy of the model. As engineers, we would like to stop sampling as soon as the learning curve starts to flatten.

Figure 4.3 is a generic algorithm that defines the process of progressive sampling. *Progressive* sampling starts by clearly defining the data used in the training set, called training pool, from which the samples would be selected (randomly, in this case) and then tested against the testing set. At each iteration, a (set of) data instance(s) of the training pool is added to the training set (Line 9). Once the data instances are selected from the training pool, they are evaluated, which in our setting means measuring the performance of the selected configuration (Line 12). The configurations and the associated performance scores are used to build the model (Line 14). The model is validated using the testing set[2], then, the accuracy is then computed. The accuracy can be quantified by any measure, such as MMRE, MBRE, Absolute Error, etc. In our setting, we assume that the measure is accuracy (higher is better). Once the accuracy score is calculated, it is compared with the accuracy score obtained before adding the new set of configurations to the training set. If the accuracy of the model (with more data) does not improve the accuracy when compared to the previous iteration (lesser data), then a life is lost. This termination criterion is widely used in the field of multi-objective optimization to determine degree of convergence [**krall2015gale**].

[2]The testing data consist of the configurations as well as the corresponding performance scores.

```
# Projective Sampling                                                    1
def projective(training, testing, thres_freq=3):                         2
  collector = list()                                                     3
  independent_vals = list()                                              4
  dependent_vals = list()                                                5
  for count in range(1, len(training)):                                  6
    # Add one configuration to the training set                          7
    independent_vals += training[count]                                  8
    # Measure the performance value for the newly                        9
    # added configuration                                                10
    dependent_vals += measure(training_set[count])                       11
    # update feature frequency table                                     12
    T = update_frequency_table(training[count])                          13
    # Build model                                                        14
    model = build_model(independent_vals, dependent_vals)                15
    # Test Model                                                         16
    perf_score = test_model(model, testing, measure(testing))           17
    # Collect the the pair of |training set|                             18
    # and performance score                                              19
    collector += [count, perf_score]                                     20
    # minimum values of the feature frequency table                      21
    if min(T) >= thresh_freq: break                                      22
  return model                                                           23
```

Figure 4.4 Pseudocode of projective sampling.

4.4.2 Projective Sampling

One of the shortcomings of progressive sampling is that the resulting performance model achieves an acceptable accuracy only after a large number of iterations, which implies high modelling cost. There is no way to actually determine the cost of modelling until the performance model is already built, which defeats its purpose, as there is a risk of over-shooting the modelling budget and still not obtain an accurate model. *Projective* sampling addresses this problem by approximating the learning curve using a minimal set of initial sampling points (configurations), thus providing the stakeholders with an estimate of the modelling cost. Sarkar et. al [33] used projective sampling to predict the number of samples required to build a performance model. The initial data points are selected by randomly adding a constant number of samples (configurations) to the training set from the training pool. In each iteration, the model is built, and the accuracy of the model is calculated using the testing data. A feature-frequency heuristic is used as the termination criterion. The feature-frequency heuristic counts the number of times a feature has been selected and deselected. Sampling stops when the counts of features selected and deselected is, at least, at a predefined threshold (*thresh_freq*).

Figure 4.4 provides a generic algorithm for projective sampling. Similar to progressive sampling, projective sampling starts with selecting samples from the training pool and adding them to the training set (Line 8). Once the samples are selected, the corresponding configurations are evaluated (Line 11). The feature-frequency table T is then updated by calculating the number of features that are selected and deselected in *independent_vals* (Line 13). The configurations

and the associated performance values are then used to build a performance model, and the accuracy is calculated (Lines 15–17). The number of configurations and the accuracy score are stored in the collector, since our objective is to estimate the learning curve. $min(T)$ holds the minimum value of the feature selection and deselection frequencies in T. Once the value of $min(T)$ is greater than *thresh_freq*, the sampled points are used to estimate the learning curve. These points are used to search for a best-fit function that can be used to extrapolate the learning curve (there are several available, including Logarithmic, Weiss and Tian, Power Law and Exponential [33]). Once the best-fit function is found, it is used to determine the point of convergence.

4.5 Rank-based approach

Typically, performance models are evaluated based on the accuracy or error. The error can be computed using[3]:

$$MMRE = \frac{\mid predicted - actual \mid}{actual} \cdot 100 \tag{4.2}$$

The key idea in this paper is to use ranking as an approach for building regression models. There are a number of advantages of using a rank-based approach:

- For the use cases listed in the introduction, *ranking is the ultimate goal*. A user may just want to identify the top-ranked configurations rather than to rank the whole space of configurations. For example, a practitioner trying to optimize an Apache Web server is searching for a set of configurations that can handle maximum load, and is not interested in the whole configuration space.

- *Ranking is extremely robust* since it is only mildly affected by errors or outliers [**kloke2012rfit**, **rosset2005ranking**]. Even though measures such as Mean Absolute Error are robust, in the configuration setting, a practitioner is often more interested in knowing the rank rather than the predicted performance scores.

- *Ranking reduces the number of training samples required to train a model*. We will demonstrate that the number of training samples required to find the optimal configuration using a rank-based approach is reduced considerably, compared to residual-based approaches which use MMRE.

It is important to note that we aim at building a performance model similar to the accurate performance model building process used by prior work as described in Section 4.4. But instead

[3]Aside: There has been a lot of criticism regarding MMRE, which shows that MMRE along with other accuracy statistics such as MMRE, MBRE has been shown to cause conclusion instability [**myrtveit2012validity**, **myrtveit2005reliability**, **foss2003simulation**].

```
# rank-based approach                                                    1
def rank_based(training, testing, lives=3):                              2
  last_score = -1                                                        3
  independent_vals = list()                                              4
  dependent_vals = list()                                                5
  for count in range(1, len(training)):                                  6
    # Add one configuration to the training set                          7
    independent_vals += training[count]                                  8
    # Measure the performance value for the newly                        9
    # added configuration                                                10
    dependent_vals += measure(training_set[count])                       11
    # Build model                                                        12
    model = build_model(independent_vals, dependent_vals)                13
    # Predicted performance values                                       14
    predicted_performance = model(testing)                               15
    # Compare the ranks of the actual performance                        16
    # scores to ranks of predicted performance scores                    17
    actual_ranks = ranks(measure(testing))                               18
    predicted_ranks = ranks(predicted_performance)                       19
    mean_RD = RD(actual_ranks, predicted_ranks)                          20
    # If current rank difference is not better than                      21
    # the previous rank difference, then loose life                      22
    if mean_rank_difference <= last_rank_difference:                     23
      lives -= 1                                                         24
    last_rank_difference = mean_RD                                       25
    # If all lives are lost, exit loop                                   26
    if lives == 0: break                                                 27
  return model                                                           28
```

Figure 4.5 Psuedocode of rank-based approach.

of using residual measures of errors, as described in Equation 4.2, which depend on residuals ($r = y - f(x)$), [4] we use a rank-based measure. While training the performance model ($f(x)$), the configuration space is iteratively sampled (from the training pool) to train the performance model. Once the model is trained, the accuracy of the model is measured by sorting the values of $y = f(x)$ from 'small' to 'large', that is:

$$f(x_1) \leq f(x_2) \leq f(x_3) \leq ... \leq f(x_n). \tag{4.3}$$

The predicted rank order is then compared to the actual rank order. The accuracy is calculated using the mean rank difference:

$$accuracy = \frac{1}{n} \cdot \sum_{i=1}^{n} \Big| rank(y_i) - rank(f(x_i)) \Big| \tag{4.4}$$

This measure simply counts how many of the pairs in the test data were ordered incorrectly by the performance model $f(x)$ and measures the average of magnitude of the ranking difference.

In Figure 4.5, we list a generic algorithm for our rank-based approach. Sampling starts by selecting samples randomly from the training pool and by adding them to the training set (Line

[4] Refer to Section 4.3 for definitions.

8). The collected sample configurations are then evaluated (Line 11). The configurations and the associated performance measure are used to build a performance model (Line 13). The generated model (CART, in our case) is used to predict the performance measure of the configurations in the testing pool (Line 16). Since the performance value of the testing pool is already measured, hence known, the ranks of the actual performance measures, and predicted performance measure are calculated. (Lines 18–19). The actual and predicted performance measure is then used to calculate the rank difference using Equation 4.4. If the rank difference of the model (with more data) does not decrease when compared to the previous generation (lesser data), then a life is lost (Lines 23–24). When all lives are expired, sampling terminates (Line 27).

The motivation behind using the parameter *lives* is: to detect convergence of the model building process. If adding more data does not improve the accuracy of the model (for example, in Figure 4.2 the accuracy of the model generated does not improve after 20 samples configuration), the sampling process should terminate to avoid resource wastage; see also Section 4.9.4.

4.6 Subject Systems

To compare residual-based approaches with our rank-based approach, we evaluate it using 21 test cases collected in 9 open-source software systems[5].

1. SS1 x264 is a video-encoding library that encodes video streams to H.264/MPEG-4 AVC format. We consider 16 features, which results in 1152 valid configurations.

2. SS2 Berkeley DB (C) is an embedded key-value-based database library that provides scalable high performance database management services to applications. We consider 18 features resulting in 2560 valid configurations.

3. SS3 SQLite is the most popular lightweight relational database management system. It is used by several browsers and operating systems as an embedded database. In our experiments, we consider 39 features that give rise to more than 3 million valid configurations.

4. SS4 wget is a software package for retrieving files using HTTP, HTTPS, and FTP. It is a non-interactive command line tool. In our experiments, we consider 16 features, which result in 188 valid configurations.

5. SS5 lrzip or Long Range ZIP is a compression program optimized for large files, consisting mainly of an extended rzip step for long distance redundant reduction and a normal compressor step. We consider 19 features, which results in 432 valid configurations.

[5]For more details on the subject systems and configurations options refer to http://tiny.cc/3wpwly

6. SS6 DUNE or the Distributed and Unified Numerics Environment, is a modular C++ library for solving partial differential equations using grid-based methods. We consider 11 feature resulting in 2305 valid configurations.

7. SS7 HSMGP or Highly Scalable MG Prototype is a prototype code for benchmarking Hierarchical Hybrid Grids data structures, algorithms, and concepts. It was designed to run on super computers. We consider 14 features resulting in 3456 valid configurations.

8. SS8 APACHE HTTP Server is a Web Server; we consider 9 features resulting in 192 valid configurations.

In addition to these 8 subject systems, we also consider Apache Storm, a distributed system, in several scenarios. The datasets were obtained from the paper by Jamshidi et al. [19]. The experiment considers three benchmarks namely:

- WordCount (WC) counts the number of occurences of the words in a text file.
- RollingSort (RS) implements a common pattern in real-time analysis that performs rolling counts of messages.
- SOL (SOL) is a network intensive topology, where the message is routed through an inter-worker network.

The experiments were conducted with all the above mentioned benchmarks on 5 cloud clusters. The experiments also contain measurement variabilities, the WC experiments were also carried out on multi-tenant cluster, which were shared with other jobs. For example, WC+RS means WC was deployed in a multi-tenant cluster with RS running on the same cluster. As a result, not only latency increased but also variability became greater. The environments considered in our experiments are:

9. SS9 WC-6D-THROUGHPUT is an environment configuration where WC is executed by varying 6 features resulting in 2879 configurations; throughput is calculated.

10. SS10 RS-6D-THROUGHPUT is an environment configuration where RS is run by varying 6 features which results in 3839 configurations; the throughput is measured.

11. SS11 WC-6D-LATENCY is an environment configuration where WC is executed by varying 6 features resulting in 2879 configurations; latency is calculated.

12. SS12 RS-6D-LATENCY is an environment configuration where RS is executed by varying 6 features, which results in 3839 configurations; latency is measured.

13. SS13 WC+RS-3D-THROUGHPUT is an environment configuration where WC is run in a multi-tenant cluster along with RS. WC is executed by varying 3 features resulting in 195 configurations; throughput is measured.

14. SS14 WC+SOL-3D-THROUGHPUT is an environment configuration where WC is run in a multi-tenant cluster along with SOL. WC is executed by varying 3 features resulting in 195 configurations; throughput is measured.

15. **SS15** WC+WC-3D-THROUGHPUT is an environment configuration where WC is run in a multi-tenant cluster along with WC. WC is executed by varying 3 features resulting in 195 configurations; throughput is measured.

16. **SS16** SOL-6D-THROUGHPUT is an environment configuration where SOL is executed by varying 6 features resulting in 2865 configurations; throughput is measured.

17. **SS17** WC-WC-3D-THROUGHPUT is an environment configuration where WC is executed by varying 3 features resulting in 755 configurations; throughput is calculated.

18. **SS18** WC+SOL-3D-LATENCY is an environment configuration where WC is run in a multi-tenant cluster along with SOL. The WC is executed by varying 3 features resulting in 195 configurations; latency is measured.

19. **SS19** WC+WC-3D-LATENCY is an environment configuration where WC is run in a multi-tenant cluster along with WC. The WC is executed by varying 3 features resulting in 195 configurations; latency is measured.

20. **SS20** SOL-6D-LATENCY is an environment configuration where SOL is executed by varying 6 features resulting in 2861 configuration setting; latency is measured.

21. **SS21** WC+RS-3D-LATENCY is an environment configuration where WC is run in a multi-tenant cluster along with RS. WC is executed by varying 3 features resulting in 195 configurations; latency is measured.

4.7 Evaluation

4.7.1 Research Questions

In the past, configuration ranking required an accurate model of the configuration space, since an inaccurate model implicitly indicates that the model has missed the trends of the configuration space. Such accurate models require the evaluation/measurement of hundreds of configuration

options for training [15, 33, 39]. There are also cases where building an accurate model is not possible, as shown in Figure 4.1 (right side).

Our research questions are geared towards finding optimal configurations when building an accurate model of a given software system is not possible. As our approach relies on ranking, our hypothesis is that we would be able to find the (near) optimal configuration using our rank-based approach while using fewer measurements, as compared to an accurate model learnt using residual-based approaches[6].

Our proposal is to embrace rank preservation but with inaccurate models and to use these models to guide configuration rankings. Therefore, to assess the feasibility and usefulness of the inaccurate model in configuration rankings, we consider the following:

- Accurate rankings found by inaccurate models using a rank-based approach, and
- the effort (number of measurements) required to build an inaccurate model.

The above considerations lead to two research questions:

RQ1*: Can inaccurate models accurately rank configurations?* Here, the optimal configurations found using an inaccurate model are compared to the more accurate models generated using residual-based approaches. The accuracy of the models is calculated using MMRE (from Equation 4.2).

RQ2*: How expensive is a rank-based approach (in terms of how many configurations must be executed)?* It is expensive to build accurate models, and our goal is to minimize the number of measurements. It is important to demonstrate that we can find optimal configurations of a system using inaccurate models as well as reducing the number of measurements.

4.7.2 Experimental Rig

For each subject system, we build a table of data, one row per valid configuration. We then run all configurations of all systems and record the performance scores (i.e., that are invoked by a benchmark). The exception is SQLite, for which we measure only 4400 configurations corresponding to feature-wise and pair-wise sampling and additionally 100 random configurations. (This is because SQLite has 3,932,160 possible configurations, which is an impractically large number of configurations to test.) To this table, we added a column showing the performance score obtained from the actual measurements for each configuration. Note that, while answering the research questions, we ensure that we never test any prediction model on the data that we used to learn the model. Next, we repeat the following procedure 20 times.

[6]Aside: It is worth keeping in mind that the approximation error in a model does not always harm. A model capable to smoothing the complex landscape of a problem can be beneficial for the search process. This sentiment has been echoed in the evolutionary algorithm literature as well [**lim2010generalizing**].

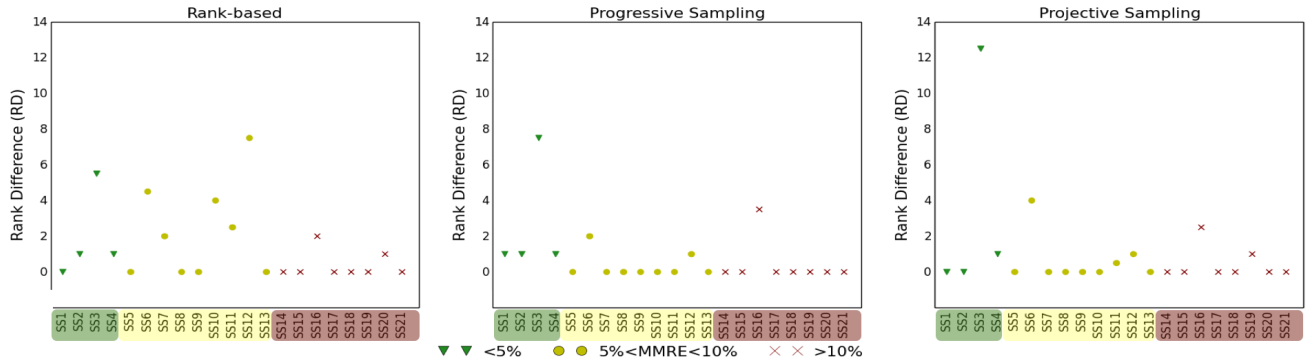

Figure 4.6 The rank difference of the prediction made by the model built using residual-based and rank-based approaches. Note that the y-axis of this chart rises to some very large values; e.g., SS3 has over three million possible configurations. Hence, the above charts could be summarised as follows: "the rank-based approach is surprisingly accurate since the rank difference is usually close to 0% of the total number of possible configurations". In this figure, ▼, ●, and × represent the subject systems (using the technique mentioned at the top of the figure), in which we could build a prediction model, where accuracy is $< 5\%$, $5\% < MMRE < 10\%$, and $> 10\%$ respectively. This is based on Figure 4.1.

To answer the research questions, we split the datasets into training pool (40%), testing pool (20%), and validation pool (40%). The experiment is conducted in the following way:

- Randomize the order of rows in the training data
- **Do**
  - Select one configuration (by sampling with replacement) and add it to the training set
  - Determine the performance scores associated with the configuration. This corresponds to a table look-up, but would entail compiling or configuring and executing a system configuration in a practical setting.
  - Using the training set and the accuracy, build a performance model using CART.
  - Using the data from the testing pool, assess the accuracy either using MMRE (as described in Equation 4.2) or the rank difference (as described in Equation 4.4).
- **While** the accuracy is greater or equal to the threshold determined by the practitioner (rank difference in the case of our rank-based approach and MMRE in the case of the residual-based approaches).

Once the model has been iteratively trained, it is used on the data in the validation pool. Please note, the learner has not been trained on the validation pool. RQ1 relates the results found by the inaccurate performance models (rank-based) to more accurate models (residual-based).

We use the absolute difference between the ranks of the configurations predicted to be the optimal configuration and the actual optimal configuration. We call this measure rank difference (RD).

$$RD = \left| rank(actual_{optimal}) - rank(predicted_{optimal}) \right| \tag{4.5}$$

Ranks are calculated by sorting the configurations based on their performance scores. The configuration with the least performance score, $rank(actual_{optimal})$, is ranked 1 and the one with highest score is ranked as N, where N is the number of configurations.

4.8 Results

4.8.1 RQ1: *Can inaccurate models accurately rank configurations?*

Rank	Treatment	Median	IQR	Median and IQR chart
SS1				
1	Projective	0.0	0.0	
1	Progressive	0.0	1.0	
2	Rank-based	2.0	8.0	
SS2				
1	Projective	0.0	1.0	
2	Rank-based	1.0	6.0	
2	Progressive	2.0	18.0	
SS3				
1	Progressive	10.0	17.0	
2	Projective	15.0	139.0	
2	Rank-based	21.0	40.0	
SS11				
1	Progressive	0.0	1.0	
2	Rank-based	1.0	3.0	
2	Projective	1.0	2.0	
SS20				
1	Projective	0.0	1.0	
1	Progressive	0.0	1.0	
2	Rank-based	5.0	19.0	

Figure 4.7 Median rank difference of 20 repeats. Median ranks is the rank difference as described in Equation 4.5, and IQR the difference between 75th percentile and 25th percentile found during multiple repeats. Lines with a dot in the middle (—●—), show the median as a round dot within the IQR. All the results are sorted by the median rank difference: a lower median value is better. The left-hand column (*Rank*) ranks the various techniques for example, when comparing various techniques for SS1, a rank-based approach has a different rank since their median rank difference is statistically different.

Figure 4.6 shows the RD of the predictions built using the rank-based approach and residual-

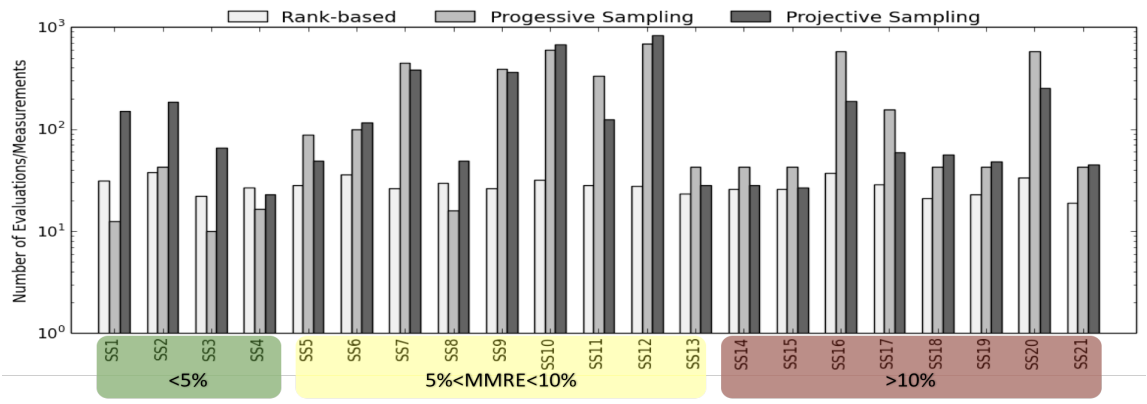

Figure 4.8 Number of measurements required to train models by different approaches. The software systems are ordered based on the accuracy scores of Figure 4.1.

based approaches [7] by learning from 40% of the training set and iteratively adding data to the training set (from the training pool), while testing against the testing set (20%). The model is then used to find the optimal configuration among the configurations in the validation dataset (40%). The horizontal axis shows subject systems. The vertical axis shows the rank difference (RD) from Equation 4.5. In this figure:

- The perfect performance model would be able to find the optimal configuration. Hence, the ideal result of this figure would be if all the points lie on the $y = 0$ or the horizontal axis. That is, the model was able to find the optimal configuration for all the subject systems ($RD = 0$).
- The markers ▼, ●, and ✕ represent the software systems where a model with a certain accuracy can be built, measured in MMRE is $< 5\%$, 5% <MMRE< 10%, and $> 10\%$ respectively.

Overall, in Figure 4.6, we find that:

- The ✕ represents software systems where the performance models are inaccurate ($> 10\%$ MMRE) and still can be used for ranking configurations, since the rank difference of these systems is always less than 4. Hence, even an inaccurate performance model can rank configurations.
- All three models built using both rank-based and residual-based approaches are able to find near optimal configurations. For example, progressive sampling for SQLite predicted the configuration whose performance score is ranked 9th in the testing set. This is good

[7] The MMRE scores for the models http://tiny.cc/bu14iy

enough since progressive sampling is able to find the 9th most performant configuration among 1861 configurations[8].

- The mean rank difference of the $predicted_{optimal}$ is 1.4, 0.77, and 0.93[9] for the rank-based approach, progressive sampling, and projective sampling respectively. Thus, a performance model can be used to rank configurations.

We claim that the rank of the optimal configuration found by the residual and rank-based approaches is the same. To verify that the similarity is statistically significant, we further studied the results using non-parametric tests, which were used by Arcuri and Briand at ICSE'11 [27]. For testing statistical significance, we used a non-parametric bootstrap test with 95% confidence [9], followed by an A12 test to check that any observed differences were not trivially small effects; that is, given two lists X and Y, count how often there are larger numbers in the former list (and if there are ties, add a half mark): $a = \forall x \in X, y \in Y \frac{\#(x>y)+0.5\cdot\#(x=y)}{|X|\cdot|Y|}$ (as per Vargha [43], we say that a "small" effect has $a < 0.6$). Lastly, to generate succinct reports, we use the Scott-Knott test to recursively divide our approaches. This recursion used A12 and bootstrapping to group together subsets that are (a) not significantly different and are (b) not just a small effect different to each other. This use of the Scott-Knott test is endorsed by Mittas and Angelis [27] and by Hassan et al. [13].

In Figure 4.7, the table shows the Scott-Knott ranks for the three approaches. The quartile charts are the Scott-Knott results for our subject systems, where the rank-based approach did not do as well as the residual-based approaches [10]. For example, the statistic test for SS1 shows that the ranks of the optimal configuration by the rank-based approach was statistically different from the ones found by the residual-based approaches. We think this is reasonably close since the median rank found by the rank-based approach is 2 out of 460 configurations, whereas residual-based approaches find the optimal configurations with a median rank of 0. As our motivation was to find optimal configurations for software systems for which performance models were difficult or infeasible to build, we look at SS20. If we look at the Skott-Knott chart for SS20, the median rank found by the rank-based approach is 5, whereas the residual-based approaches could find the optimal configurations very consistently (IQR=1). But as engineers, we feel that this is close because we are able to find the 5th best configuration using 33 measurements compared to 251 and 576 measurements used for progressive and projective sampling, respectively. Overall, our results indicate that:

[8]Since we test only on 40% of the possible configuration space (40% of 4653).
[9]The median rank difference is 0 for all the approaches.
[10]For complete Skott-Knott charts, refer to http://geekpic.net/pm-1GUTPZ.html.

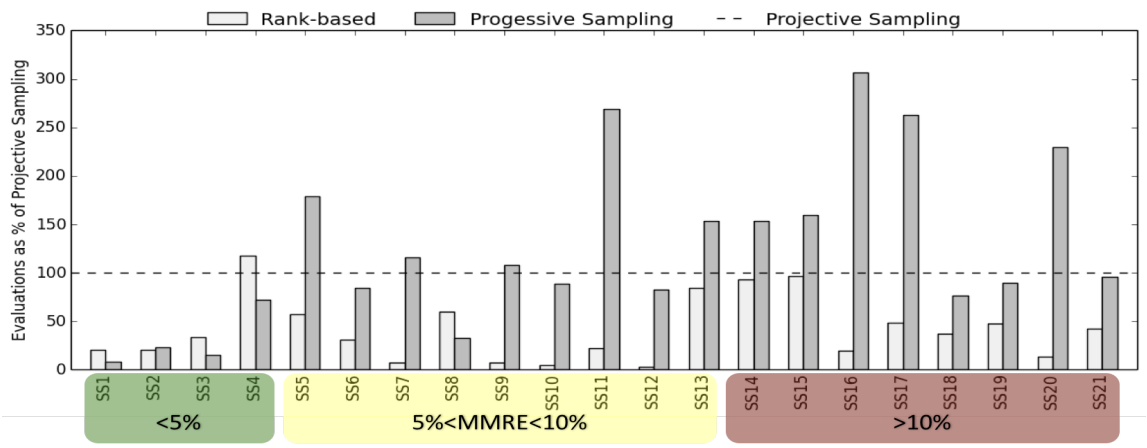

Figure 4.9 The percentage of measurement used for training models with respect to the number of measurements used by projective sampling (dashed line). The rank-based approach uses almost 10 times less measurements than the residual-based approaches. The subject systems are ordered based on the accuracy scores of Figure 4.1.

A rank preserving (probably inaccurate) model can be useful in finding (near) optimal configurations of a software system using a rank-based approach.

4.8.2 RQ2: *How expensive is a rank-based approach?*

To answer the question of whether we can recommend the rank-based approach as a cheap method for finding the optimal configuration, it is important to demonstrate that rank-based models are indeed cheap to build. In our setting, the cost of a model is determined by the number of measurements required to train the model. Figure 4.8 demonstrates this relationship. The vertical axis denotes the number of measurements in log scale and horizontal axis represents the subject systems.

In the systems SS1–SS4 (green band), the number of measurements required by the rank-based approach is less than for projective sampling and more than for progressive sampling. This is because the subject systems are easy to model. For the systems SS4–SS13 (yellow band), the number of measurements required to build models using the rank-based approach is less than residual-based approaches, with the exception of SS8. Note that, as building accurate models becomes difficult, the difference between the number of measurements required by the rank-based approach and residual-based approaches increases. For the systems SS14–SS21 (red band), the number of measurements required by the rank-based approach to build a model is always less than for residual-based approaches, with significant gains for SS19–SS21.

In Figure 4.9, the ratio of the measurements of different approaches are represented as the

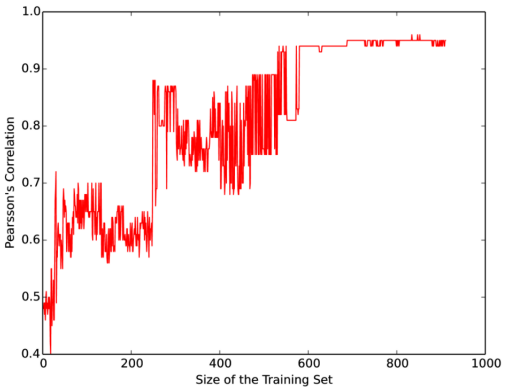

Figure 4.10 The correlation between actual and predicted performance values (not ranks) increases as the training set increases. This is evidence that the model does learn as training progresses.

percentage of number of measurements required by projective sampling – since it uses the most measurements in 50% of the subject systems. For example, in SS5, the number of measurements used by progressive sampling is twice as much as used by projective sampling, whereas the rank-based approach uses half of the total number of measurements used by projective sampling. We observe that the number of measurements required by the rank-based approach is much lower than for the residual-based approaches, with the only exceptions of SS4 and SS8. We argue that such outliers are not of a big concern since the motivation of rank-based approach is to find optimal configurations for software systems, where an accurate model is infeasible.

To summarize, the number of samples required by the rank-based approach is much smaller than for residual-based approaches. There are $\frac{4}{21}$ cases where residual-based approaches (progressive sampling) use fewer measurements. The subject systems where residual-based approaches use fewer measurements are systems where accurate models are easier to build (green and yellow band):

> Models built using the rank-based approach require fewer measurements than residual-based approaches. In $\frac{8}{21}$ of the cases, the number of measurements is an order of magnitude smaller than residual-based approaches.

4.9 Discussion

4.9.1 How is rank difference useful in configuration optimization?

The objective of the modelling task considered here is to assist practitioners to find optimal configuration/s. We use a performance model, like traditional approaches, to solve this problem,

but rather than using a residual-based accuracy measure, we use rank difference. Rank difference can also be thought as a correlation-based metric, since rank difference essentially tries to preserve the ordering of performance scores.

In our rank-based approach, we do not train the model to predict the dependent values of the testing set. But rather, we attempt to train the model to predict the dependent value that is correlated to the actual values of the testing set. So, during iterative sampling based on the rank-based approach, we should see an increase in the correlation coefficient as the training progresses. Figure 4.10 shows how the correlation between actual and the predicted values increases as the size of the training set grows. From the combination of Figure 4.1 and Figure 4.6, we see that even an inaccurate model can be used to find an optimal configuration.[11]

4.9.2 Can inaccurate models be built using residual-based approaches?

We have already shown that a rank preserving (probably inaccurate) model is sufficient to find the optimal configuration of a given system. MMRE can be used as a stopping criterion, but as we have seen with residual-based approaches, they require a larger training set and hence are not cost effective. This is because, with residual-based approaches, unlike our rank-based approach, it is not possible to know when to terminate sampling. It may be noted that rank difference can be easily replaced with a correlation-based approach such as Pearson's or Spearman's correlation.

4.9.3 Can we predict the complexity of a system to determine which approach to use?

From our results, we observe that a rank-based approach is not as effective as the residual-based approaches for software systems that can be modelled accurately (green band). Hence, it is important to distinguish between software systems where the rank-based approach is suitable and software systems where residual-based approaches are suitable. This is relatively straight-forward since both rank-based and residual-based approaches use random sampling to select the samples. The primary difference between the approaches is the termination criterion. The rank-based approach uses rank difference as the termination criterion whereas residual-based approaches use criterion based on MMRE, etc. Hence, it is possible to use both techniques simultaneously. If a practitioner observes that the accuracy of the model during the building process is high (as in case of SS2), residual-based approaches would be preferred. Conversely, if the accuracy of the model is low (as in the case of , SS21), the rank-based approach would be preferred.

[11] This also shows how a correlation-based measure can be used as a stopping criterion.

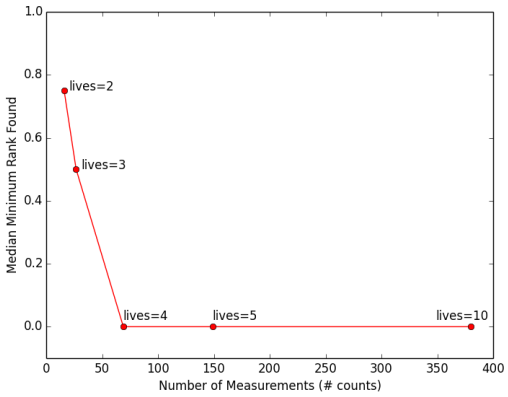

Figure 4.11 The trade-off between the number of measurements or size of the training set and the number of *lives*.

4.9.4 What is the trade-off between the number of lives and the number of measurements?

Our rank-based approach requires that the practitioner defines a termination criterion (*lives* in our setting) before the sampling process commences, which is a similar to progressive sampling. The termination criterion preempts the process of model building based on an accuracy measure. The rank-based approach uses rank difference as the termination criterion, whereas residual-based approaches use residual-based measures. In our experiments, the number of measurements or the size of the training set depends on the termination criterion (*lives*). An early termination of the sampling process would lead to a sub-optimal configuration, while late termination would result in resource wastage. Hence, it is important to discuss the trade-off between the number of *lives* and the number of measurements. In Figure 4.11, we show the trade-off between the median minimum ranks found and the number of measurements (size of training set). The markers of the figure are annotated with the values of *lives*. The trade-off characterizes the relationship between two conflicting objectives, for example, point (*lives*=2) requires very few measurements but the minimum rank found is the highest, whereas point (*lives*=10) requires a large number of measurements but is able to find the best performing configuration. Note, this curve is an aggregate of the trade-off curves of all the software systems discussed in this paper[12]. Since our objective is to minimize the number of measurements while reducing rank difference, we assign the value of 3 to *lives* for the purposes of our experiments.

[12]Complete trade-off curves can be found at http://tiny.cc/kgs2iy or http://tiny.cc/rank-param.

4.10 Reliability and Validity

Reliability refers to the consistency of the results obtained from the research. For example, how well can independent researchers reproduce the study? To increase external reliability, we took care to either clearly define our algorithms or use implementations from the public domain (scikit-learn) [35]. Also, all data and code used in this work are available on-line.[13]

Validity refers to the extent to which a piece of research actually investigates what the researcher purports to investigate [38]. *Internal validity* is concerned with whether the differences found in the treatments can be ascribed to the treatments under study.

For SQLite, we cannot measure all possible configurations in reasonable time. Hence, we sampled only 100 configurations to compare prediction and actual values. We are aware that this evaluation leaves room for outliers and that measurement bias can cause false interpretations [26]. Since we limit our attention to predicting performance for a given workload, we did not vary benchmarks.

We aimed at increasing *external validity* by choosing subject systems from different domains with different configuration mechanisms. Furthermore, our subject systems are deployed and used in the real world.

4.11 Conclusion

Configurable systems are widely used in practice, but it requires careful tuning to adapt them to a particular setting. State-of-the-art approaches use a residual-based technique to guide the search for optimal configurations. The model-building process involves iterative sampling used along with a residual-based accuracy measure to determine the termination point. These approaches require too many measurements and hence are expensive. To overcome the requirement of a highly accurate model, we propose a rank-based approach, which requires a lesser number of measurements and finds the optimal configuration just by using the ranks of configurations as an evaluation criterion.

Our key findings are the following. First, a highly accurate model is not required for configuration optimization of a software system. We demonstrated how a rank-preserving (possibly even inaccurate) model can still be useful for ranking configurations, whereas a model with accuracy as low as 26% can be useful for configuration ranking. Second, we introduce a new rank-based approach that can be used to decide when to stop iterative sampling. We show how a rank-based approach is not trained to predict the raw performance score but rather learns the model, so that the predicted values are correlated to actual performance scores.

[13] https://github.com/ai-se/Reimplement/tree/cleaned_version

To compare the rank-based approach to the state-of-the-art residual-based approaches (projective and progressive sampling), we conducted a number of experiments on 9 real-world configurable systems to demonstrate the effectiveness of our approach. We observed that the rank-based approach is effective to find the optimal configurations for most subject systems while using fewer measurements than residual-based approaches. The only exceptions are subject systems for which building an accurate model is easy, anyway.

Chapter 5

Finding faster configurations using flash

This chapter presents an approach to estimating end-to-end performance of distributed storage systems. We explain why low-level performance metrics are a desirable proxy for estimating end-to-end performance. We then present our automatic model building tool for generating robust and accurate performance models.

Chapter 6

Conclusions and Future Work

REFERENCES

[1] Bettenburg, N. et al. "Think locally, act globally: Improving defect and effort prediction models". *Proceedings of the 9th IEEE Working Conference on Mining Software Repositories*. IEEE Press. 2012, pp. 60–69.

[2] Bettenburg, N. et al. "Towards improving statistical modeling of software engineering data: think locally, act globally!" *Empirical Software Engineering* **20**.2 (2015), pp. 294–335.

[3] Bingham, E. & Mannila, H. "Random projection in dimensionality reduction: applications to image and text data". *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2001, pp. 245–250.

[4] Boley, D. "Principal direction divisive partitioning". *Data mining and knowledge discovery* **2**.4 (1998), pp. 325–344.

[5] Breiman, L. et al. *Classification and regression trees*. CRC press, 1984.

[6] Deb, K. et al. "A fast and elitist multiobjective genetic algorithm: NSGA-II". *IEEE Transactions on Evolutionary Computation* **6**.2 (2002), pp. 182–197.

[7] Deiters, C. et al. "Using spectral clustering to automate identification and optimization of component structures". *Realizing Artificial Intelligence Synergies in Software Engineering (RAISE), 2013 2nd International Workshop on*. IEEE. 2013, pp. 14–20.

[8] Du, Q. & Fowler, J. E. "Low-complexity principal component analysis for hyperspectral image compression". *International Journal of High Performance Computing Applications* **22**.4 (2008), pp. 438–448.

[9] Efron, B. & Tibshirani, R. J. *An introduction to the bootstrap*. CRC, 1993.

[10] Faloutsos, C. & Lin, K.-I. "FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets". Vol. 24. 2. ACM, 1995.

[11] Faloutsos, C. & Lin, K.-I. *FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets*. Vol. 24. 2. ACM, 1995.

[12] Fletcher, R. *Practical methods of optimization*. John Wiley & Sons, 2013.

[13] Ghotra, B. et al. "Revisiting the Impact of Classification Techniques on the Performance of Defect Prediction Models". *37th IEEE International Conference on Software Engineering*. 2015.

[14] Grassberger, P. & Procaccia, I. "Measuring the strangeness of strange attractors". *The Theory of Chaotic Attractors*. Springer, 2004, pp. 170–189.

[15] Guo, J. et al. "Variability-aware performance prediction: A statistical learning approach". *IEEE/ACM 28th International Conference on Automated Software Engineering*. IEEE. 2013, pp. 301–311.

[16] Hamerly, G. "Making k-means Even Faster." Society for Industrial and Applied Mathematics.

[17] Harman, M. et al. "Search-based software engineering: Trends, techniques and applications". *ACM Computing Surveys* **45**.1 (2012), p. 11.

[18] Ilin, A. & Raiko, T. "Practical approaches to principal component analysis in the presence of missing values". *The Journal of Machine Learning Research* **11** (2010), pp. 1957–2000.

[19] Jamshidi, P. & Casale, G. "An Uncertainty-Aware Approach to Optimal Configuration of Stream Processing Systems". *2016 IEEE 24th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS)*. 2016, pp. 39–48.

[20] Jolliffe, I. *Principal component analysis*. Wiley Online Library, 2002.

[21] Kamvar, K. et al. "Spectral learning". *International Joint Conference of Artificial Intelligence*. Stanford InfoLab. 2003.

[22] Krall, J. et al. "GALE: Geometric Active Learning for Search-Based Software Engineering". *IEEE Transactions on Software Engineering* **41**.10 (2015), pp. 1001–1018.

[23] Kuhn, D. R. et al. *Introduction to combinatorial testing*. CRC press, 2013.

[24] Loshchilov, I. G. "Surrogate-assisted evolutionary algorithms". PhD thesis. 2013.

[25] Marker, B. et al. "Understanding performance stairs: Elucidating heuristics". *Proceedings of the 29th ACM/IEEE international conference on Automated software engineering*. ACM. 2014, pp. 301–312.

[26] Menzies, T. et al. "Local versus global lessons for defect prediction and effort estimation". *IEEE Transactions on Software Engineering* **39**.6 (2013), pp. 822–834.

[27] Mittas, N. & Angelis, L. "Ranking and Clustering Software Cost Estimation Models through a Multiple Comparisons Algorithm". *IEEE Transactions of Software Engineering* (2013).

[28] Nair, V. et al. "Faster discovery of faster system configurations with spectral learning". *Automated Software Engineering* (2017), pp. 1–31.

[29] Nair, V. et al. "Using bad learners to find good configurations". *arXiv preprint arXiv:1702.05701* (2017).

[30] Oh, J. et al. “Finding Near-optimal Configurations in Product Lines by Random Sampling”. *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. ESEC/FSE 2017. Paderborn, Germany: ACM, 2017, pp. 61–71.

[31] Platt, J. “FastMap, MetricMap, and Landmark MDS are all Nystrom Algorithms”. Ed. by Cowell, R. G. & Ghahramani, Z. Society for Artificial Intelligence and Statistics, 2005, pp. 261–268.

[32] Pukelsheim, F. *Optimal Design of Experiments*. Vol. 50. Society for Industrial and Applied Mathematics, 1993.

[33] Sarkar, A. et al. “Cost-Efficient Sampling for Performance Prediction of Configurable Systems”. *30th IEEE/ACM International Conference on Automated Software Engineering*. IEEE. 2015, pp. 342–352.

[34] Sayyad, A. S. et al. “Scalable product line configuration: A straw to break the camel’s back”. *IEEE/ACM 28th International Conference on Automated Software Engineering*. IEEE. 2013, pp. 465–474.

[35] *scikit-learn*. http://scikit-learn.org.

[36] Shi, J. & Malik, J. “Normalized cuts and image segmentation”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**.8 (2000), pp. 888–905.

[37] Shi, J. & Malik, J. “Normalized cuts and image segmentation”. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **22**.8 (2000), pp. 888–905.

[38] Siegmund, J. et al. “Views on internal and external validity in empirical software engineering”. *Proceedings of the 37th International Conference on Software Engineering*. IEEE. 2015, pp. 9–19.

[39] Siegmund, N. et al. “Predicting performance via automated feature-interaction detection”. *Proceedings of the 34th International Conference on Software Engineering*. IEEE Press. 2012, pp. 167–177.

[40] Siegmund, N. et al. “Performance-influence models for highly configurable systems”. *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*. ACM. 2015, pp. 284–294.

[41] Storn, R. & Price, K. “Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces”. *Journal of global optimization* **11**.4 (1997), pp. 341–359.

[42] Theisen, C. et al. “Approximating attack surfaces with stack traces”. *Proceedings of the 37th International Conference on Software Engineering*. IEEE Press. 2015, pp. 199–208.

[43] Vargha, A. & Delaney, H. D. “A Critique and Improvement of the CL Common Language Effect Size Statistics of McGraw and Wong”. *Journal of Educational and Behavioral Statistics* (2000).

[44] Wang, Z. et al. “Bayesian optimization in a billion dimensions via random embeddings”. *Journal of Artificial Intelligence Research* **55** (2016), pp. 361–387.

[45] Weiss, G. M. & Tian, Y. “Maximizing classifier utility when there are data acquisition and modeling costs”. *Data Mining and Knowledge Discovery* **17**.2 (2008), pp. 253–282.

[46] Xu, T. et al. “Hey, you have given me too many knobs!: Understanding and dealing with over-designed configuration in system software”. *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*. ACM. 2015, pp. 307–319.

[47] Zhang, F. et al. “Cross-project Defect Prediction Using A Connectivity-based Unsupervised Classifier”. *Proceedings of the 38th International Conference on Software Engineering (ICSE'16)*. 2016.

[48] Zhang, Y. et al. “Performance Prediction of Configurable Software Systems by Fourier Learning”. *30th IEEE/ACM International Conference on Automated Software Engineering*. IEEE. 2015, pp. 365–373.

[49] Zuluaga, M. et al. “Active learning for multi-objective optimization”. *Proceedings of the 30th International Conference on Machine Learning*. 2013, pp. 462–470.

[50] Zuluaga, M. et al. “ε-PAL: An Active Learning Approach to the Multi-Objective Optimization Problem”. *The Journal of Machine Learning Research* **17**.1 (2016), pp. 3619–3650.