ABSTRACT

Title of Thesis: EFFECTS OF DIVERSE INITIALIZATION ON BAYESIAN OPTIMIZERS

EESH KAMRAH
Master of Science, 2023

Thesis Directed by: Professor MARK FUGE
Department of Mechanical Engineering

Design researchers have struggled to produce quantitative predictions for exactly why and when diversity might help or hinder design search efforts. This thesis addresses that problem by studying one ubiquitously used search strategy—Bayesian Optimization (BO)—on different ND test problems with modifiable convexity and difficulty. Specifically, we test how providing diverse versus non-diverse initial samples to BO affects its performance during search and introduce a fast ranked-DPP method for computing diverse sets, which we need to detect sets of highly diverse or non-diverse initial samples.

We initially found, to our surprise, that diversity did not appear to affect BO, neither helping nor hurting the optimizer's convergence. However, follow-on experiments illuminated a key trade-off. Non-diverse initial samples hastened posterior convergence for the underlying model hyper-parameters—a *Model Building* advantage. In contrast, diverse initial samples accelerated exploring the function itself—a *Space Exploration* advantage. Both advantages help BO, but

in different ways, and the initial sample diversity directly modulates how BO trades those advantages. Indeed, we show that fixing the BO hyper-parameters removes the Model Building advantage, causing diverse initial samples to always outperform models trained with non-diverse samples. These findings shed light on why, at least for BO-type optimizers, the use of diversity has mixed effects and cautions against the ubiquitous use of space-filling initializations in BO. To the extent that humans use explore-exploit search strategies similar to BO, our results provide a testable conjecture for why and when diversity may affect human-subject or design team experiments.

The thesis is organized as follows: Chapter 2 provides an overview of existing studies that explore the impact of different initial stimuli. In Chapter 3, we explain the methodology used in the subsequent experiments. Chapter 4 presents the results of our initial study on the diverse initialization of BO (Bayesian Optimization) applied to the wildcat wells function. In this chapter we also investigate the conditions under which less diverse initial examples perform better and expand on these findings in Chapter 5 by considering additional ND continuous functions. The final chapter discusses the limitations of our findings and proposes potential areas for future research.

EFFECTS OF DIVERSE INITIALIZATION ON BAYESIAN OPTIMIZERS

by

Eesh Kamrah

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master's of Science
2023

Advisory Committee:
Professor Mark Fuge, Chair/Advisor
Professor Shapour Azram
Professor Nikhil Chopra

Dedication

I dedicate this thesis to my late Grandfather, Sh. Krishan Kumar Puri, who I lost during the pandemic. He was not just a grandfather to me but also a father figure and a mentor who played a pivotal role in shaping my life and career.

As a child, he raised me with love and discipline, instilling in me the values of hard work, perseverance, and dedication. He was strict, but it was always for my benefit. His unwavering support and guidance inspired me to pursue my dreams and go into academia, just like he did as an instructor himself.

Despite the distance and time difference, he always made it a point to check up on me regularly and encourage me in my endeavors. His words of wisdom, his warmth, and his unwavering love were a constant source of motivation for me, even as I pursued my graduate studies in the United States.

The pandemic brought with it unimaginable loss and separation, and I deeply regret not being able to be there with him in his final moments. But, as I write these words, I know that he is looking down on me with pride and love. This thesis is a testament to his influence in my life and a small way of honoring his memory.

Grandpa, this is for you. I will always be grateful for everything you did for me and the impact you had on my life.

Acknowledgments

I would like to express my sincere gratitude to everyone who has supported me in the completion of my thesis. Firstly, I would like to thank my advisor Dr. Mark Fuge for his invaluable guidance, encouragement, and support throughout my entire graduate program. I am extremely grateful to have had such an amazing mentor, and if I had to start all over again, I would always join the IDEAL lab. Mark has been an integral part of my research journey, and I couldn't have done it without him.

I would also like to thank Dr. Ghoreishi for her assistance and support during the initial stages of my research on the Diversity Solution project. Her valuable insights and suggestions were instrumental in shaping the direction of my research.

I am grateful to my lab mates Alec, Milad, Nicholas, and Quiyi for their constant support, encouragement, and insightful discussions over the past two years. Their contributions have been invaluable, and I am fortunate to have worked with such a talented and dedicated team.

I would like to express my gratitude to my friends, Vaibhav, Rohan, Prachi, Aditya, Julie, Sameer, Raghav, Kabir, Tanmay, and Abhinav, who have been a source of motivation and positivity throughout my graduate program. I apologize if I have missed anyone, but please know that your support has been appreciated.

I would like to acknowledge my parents and siblings for their unwavering support and encouragement throughout my academic journey. I would also like to extend my gratitude to my

family in the USA, who have been a constant source of support during a difficult time in my life.

Finally, I would like to thank the National Science Foundation through award #1826083 for their financial support, which has enabled me to complete this research.

Once again, thank you to everyone who has contributed to the successful completion of my thesis.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BO	Bayesian Optimzation
DDG	Delta Design Game
DbA	Design-by-Analogy
DPP	Determinatal Point Processes
LHS	Latin Hypercube Sampling
MTURK	Amazon Mechanical Turk
ND	N - Dimensional

Chapter 1: Introduction

Design in engineering is a complex task and often requires some examples or analogies during the design process. This is in-fact a well studied field within design engineering, DbA (Design-by-Analogy). Examples/analogies have shown to be helpful to designers during the ideation process and help them come up with innovative solutions [1, 2]. For example, a team of engineers wanted to soften the sonic boom effect that high-speed trains face. A solution came from an analogy in nature: a kingfisher can slice through the air at high speeds to catch its prey. Engineers used this and constructed the train's front end to imitate a kingfisher [3].

This work looks at when and how providing diverse versus non-diverse stimuli affects the solution quality of a design search process. There has been past research, particularly in the field of design cognition, that has shown how providing diverse examples helps produce better results [1, 4]. These studies are excellent examples of showing how initial solutions help and hinder design processes. The current literature contains a wide range of results regarding the impact of diversity on the outcome of a problem, making it difficult to accurately predict whether it will be helpful for a new problem or individual. There are several examples that demonstrate the positive effects of diversity on novelty and the diversity of ideas [5, 6, 7], but substantially more mixed results on the effects of diversity on solution *quality*, with some observations of positive effects [8, 9, 10], some null or contingent effects [6, 11, 12, 13, 14, 15, 16], and even

some negative effects on solution quality [17, 18].

Diversity is an important consideration in optimization research, and there are various strategies for incorporating it into search algorithms. Common approaches include initializing algorithms with different strategies, such as Latin Hypercube Sampling (LHS)[19] and quasi-random methods[20, 21, 22], which aim to uniformly cover the search space [23]. In addition, some meta-heuristic optimizers [24] like Particle Swarm Optimization (PSO), Simulated Annealing (SA), and Genetic Algorithms (GA) incorporate diversity-encouraging loss functions into their core search algorithms, with NSGA-II [25] being one of the most well-known diversity-inducing ones. For Bayesian optimization (BO), diversity can be built directly into the acquisition function used to sample new points from the Gaussian Process posterior [26]. However, the effect of diversity on optimization performance is often problem-specific and challenging to predict apriori [24]. While encouraging initial diversity is generally thought to improve convergence speed and find better global optima, it is important to consider when and why diversity might hurt rather than help our search for good designs.

In the context of this thesis, quantifying the diversity of a set of samples is a crucial but challenging task. Common approaches involve using hyper-volume maximizing methods that can maximize coverage over a fixed amount of points. However, these methods become computationally expensive as they scale combinatorially. To address this challenge, researchers have used Determinantal Point Processes (DPPs) [27] to enable fast polynomial time approximations of diverse sets. Among the different types of DPPs, k-DPPs are a popular tool that enables sampling diverse sets with polynomial complexity [28]. Thus, researchers can sample diverse sets with relative ease, which is essential for optimization and design engineering applications. Despite the usefulness of k-DPPs, there are still limitations to consider. For instance, sampling

percentile sets from the DPP distribution to obtain the top 5%, median, or lowest 5% of diverse sets can become exceedingly slow when dealing with a large sample pool. Therefore, it is crucial to develop a new approach to allow us to quantify diversity in optimization and design engineering.

Building upon this approach, this thesis investigates the effect of diverse initialization on BO (Bayesian Optimization). We initialize a global optimization routine on a ND continuous landscape generated using a test function generator that was developed for this study. While undertaking this study we learned that both less-diverse and diverse samples provide different advantages to the search process. The less-diverse samples gain an advantage with their ability to study the local landscape, we call this the 'Model Building Advantage', while the diverse designs experience a 'Space Exploration' advantage. In the following sections I will provide a more detailed overview of the structure of this thesis, highlighting the key contributions of each chapter:

Summary of Chapter 2: This chapter answers 4 important questions: Why should one model design search as Bayesian optimization? Why do we need a modular benchmark function generator for this study? What does it even mean for samples to be diverse? And how does diversity in initial inputs affect optimizers? It does so by reviewing literature in the respective areas and identifying the need for the specific contribution.

Summary of Chapter 3: This chapter presents the methodology required to understand the experimental setup found in the studies in Chapters 4 and 5. The chapter begins with an extensive background on Bayesian Optimization (BO), followed by a discussion on the specific methodol-

ogy employed for the utilization of BO in this design study. Next, the details of the wildcat-wells test function generator are explored, encompassing its design and implementation. Lastly, the chapter elucidates the approach employed for diverse sampling and also presents how the upper error bound for the diverse sampling method is constructed.

Summary of Chapter 4: In this chapter, we look at the results from our initial study on the effects of diversity on the performance of BO for the wildcat wells surface. The chapter provides insights about our surprising findings that non-diverse examples outperform diverse examples in several cases. We then try to answer the question of why the low-diversity examples outperform the diverse examples, given the historical precedent for deferring to space-filling initialization. We do so by taking on two studies: the first confirms our hypothesized advantage for low-diversity samples in terms of faster posterior convergence for the BO kernel. We show that this is the causal effect of the improved performance through an ablation study that fixes the kernel to a known ground truth, confirming our hypothesis.

Summary of Chapter 5: In this last study, we extend the results from wildcat wells to three other widely used test functions —namely, the Sphere, Rosenbrock, and Rastrigin functions. We also test how the effect changes as a function of the problem dimension.

Finally, in Chapter 6, we summarize our findings, discuss the implications of our results, and suggest areas for future research. Chapter 6 also explores the limitations of this thesis and how these limitations can be addressed, as well as future research directions for this work. Portions of the work presented in this thesis were also published as an archival paper in [29].

Chapter 2: Background and Related Work

Before describing our particular experiment and results, we will first review why BO is a meaningful and generalizable class of search algorithm to use, as well as past work that has tried to understand how diversity affects search processes such as optimization.

2.1 Why model design search as Bayesian optimization?

While this thesis addresses only BO, this is an important algorithm in that it plays an outsized role within the design research and optimization community. For example, BO underlies a vast number of industrially-relevant gradient-free surrogate modeling approaches implemented in major design or analysis packages, where it is referred to under a variety of names, including Kriging methods or meta-modeling [30, 31]. Its use in applications of computationally expensive multidisciplinary optimization problems is, while not unilateral [32], quite widespread. Likewise, researchers studying human designers often use BO as a proxy model [33] to understand human search, due to the interplay between exploration and exploitation that lies at the heart of most BO acquisition functions like Expected Improvement. More generally, there is a robust history of fruitful research in cognitive science modeling human cognition as Bayesian processing [34], such as concept learning in cognitive development [35], causal learning [36], and analogical reasoning [37].

Even though, studies in areas like airfoil and material design optimization have highlighted the impact of selecting the correct initial designs with BO on reducing run-time for experiments [38, 39], the bulk of BO-related papers focus on new algorithms or acquisition functions, few papers focus on how BO is initialized, preferring instead the general use of space-filling initializations that have a long history in the field of Optimal Experiment Design [32]. In contrast, this thesis shows that in certain situations that faith in space-filling designs might be misplaced, particularly when the BO kernel hyper-parameters are adjusted or fit during search.

2.2 Why do we need a modular benchmark function generator?

To rigorously test our hypothesis and to generalize the solution it was important to test our hypothesis against problems of varying complexity and several problems of similar complexity. However, most benchmark functions are not modular and do not allow for generating surfaces of varying complexities or similar complexity, limiting their usefulness for algorithm comparison and analysis. In recent years, there has been a growing interest in developing modular benchmark functions that can generate surfaces of varying complexities with different seeds in dynamic optimization problems, as noted in recent surveys [40, 41]. For example, the work by Ali Ahrari et al. [42] generates benchmark functions by composing a set of fixed functions, there is a parameter that can be used to control the complexity of the problem, but the generator lacks the ability to generate functions of similar complexity because the generator is deterministic.

Our proposed function generator, wildcat wells, which we detail in Chapter 3, addresses all the mentioned concerns by generating test functions using a mixture of simplex noise and multivariate normal distribution. This provides a flexible way to generate surfaces of varying

complexities as well as similar complexity. We present the details of the wildcat-wells test function generator, including its design and implementation in Section 3.1. In addition, we show in Chapter 5 how our primary results vary as we change the target optimization function.

2.3 What does it even mean for samples to be diverse?

As a practical matter, if we wish to study how diverse samples impact BO, we face a subtle but surprisingly non-trivial problem: how exactly do you quantify whether one set of samples is more or less diverse than another? This is a set-based (*i.e.*, combinatorially large) problem with its own rich history too large to cover extensively here, however our past work on diversity measurement [43, 44, 45], computation [46], and optimization [47] provides further pointers for interested readers, and in particular the thesis of Ahmed provides a good starting point for the broader literature and background in this area [48].

For the purposes of understanding how this thesis relates to existing approaches, it suffices to know the following regarding common approaches to quantifying diversity: (1) most diversity measurement approaches focus on some variant of a hyper-volume objective spanned by the set of selected points; (2) since this measure depends on *a set* rather than individual points, it becomes combinatorially expensive, necessitating fast polynomial-time approximation, one common tool for which is a Determinantal Point Process (DPP) [27]; however, (3) while sampling the most diverse set via DPPs is easy, sampling percentile sets from the DPP distribution to get the top 5%, median, or lowest 5% of diverse sets becomes exceedingly slow for a large sample pool.

In contrast, for this thesis, we created a faster DPP-type sampling method to extract different percentiles of the distribution without actually needing to observe the entire DPP distri-

bution and whose sampling error we can bound using concentration inequalities. Section 3.4 provides further mathematical background, including information on DPP hyper-parameters and how to select them intelligently, and the Supplemental Material provides further algorithmic details. With an understanding of diversity distribution measures in hand, we can now address diversity's specific effects on optimization more generally.

How does diversity in initial inputs affect optimizers? While there are a number of papers that propose either different initialization strategies or benchmarking of existing strategies for optimization, there is limited prior work addressing the direct effect of initial sample diversity.

For general reviews and benchmarking on how to initialize optimizers and the effects of different strategies, papers such as [21, 24] compare initialization strategies for particular optimizers and quantify performance differences. An overall observation across these contributions is the inability of a single initialization method to improve performance across functions of varying complexity. These studies also do not directly measure or address the role of sample diversity directly, only noting such behavior as it correlates indirectly with the sampling strategy.

A second body of work tries to customize initialization strategies on a per-problem basis, often achieving faster convergence on domain-specific problems [20, 23, 49, 50, 51]. While useful in their designed domain, these studies do not directly address the role of diversity either. In contrast, this thesis addresses diversity directly using properties of BO that are sufficiently general to apply across multiple domains and applications.

Lastly, how to initialize optimizers has garnered new interest from the machine learning community, for example in the initial settings of weights and biases in a Neural Network and the downstream effects on network performance [52, 53]. There is also general interest in how

to collect diverse samples during learning, either in an Active Learning [54] or Reinforcement Learning context [55, 56]; however, those lines of work address only diversity throughout data collection, rather than the impact of initial samples considered in this thesis.

2.4 What does this thesis contribute beyond past work?

This thesis's specific contributions are:

1. To compute diversity: we describe a fast DPP-based diversity scoring method for selecting diverse initial examples with a fixed size k. Any set of size k with these initial examples can be then used to approximate the percentile of diversity that the set belongs to. This method requires selecting a hyper-parameter relating to the DPP measure. We describe a principled method for selecting this parameter in Section 2, and provide numerical evidence of the improved sampling performance in the Sec. 3.4. Compared to prior work, this makes percentile sampling of DPP distributions computationally tractable.

2. To study effects on BO: we empirically evaluate how diverse initial samples affect the convergence rate of a Bayesian Optimizers on ND continuous problems. Section 4.2 finds that low diversity samples provide a *Model Building* advantage to BO while diverse samples provide a *Space Exploration* advantage that helps BO converge faster. Section 4.4 shows that removing the model building advantage makes having diverse initial samples uniformly better than non-diverse samples.[1]

The next chapter describes our overall experimental approach and common procedures used

[1]For grammatical simplicity and narrative flow, we will use the phrase "non-diverse" throughout the thesis to refer to cases where samples are taken from the 5th percentile of diverse sets—these are technically "low-diversity" rather than being absolutely "non-diverse" which would occur when all points in the set are identical, but we trust that readers can keep this minor semantic distinction in mind.

across all three of our main experiments. This will also introduce our diverse sampling method as well as wildcat wells function generator.

Chapter 3: Experiment Methodology

This chapter describes the main methodology and choices behind how we chose to study the effect of diversity on Bayesian Optimization. This includes (1) our choice of test function generator, (2) how we set up the Bayesian Optimization (BO), (3) how we determined the ground truth BO kernel parameters, and (4) how we sampled diverse sets.

3.1 Methodology wildcat wells

Our function generator derives its idea from a static test function used in Mason et al. [57]. To modulate the functions we used four factors, which are described below:

1. The *ruggedness amplitude*, in the range of [0,1], controls the relative height of noise added to the search environment, compared to the height of the peaks. Increasing this parameter makes the noise being added more influential. Setting this parameter to 1 at a low smoothness would give the wildcatwells function infinite peaks.

2. The *smoothness*, in the range of [0,1], controls the degree of local correlation of X in the grid. Intuitively, if smoothness was high (closer to 1), then the surface would mean that the points around each other are of relatively the same value. Thus, gradient based optimizer should benefit directly from increasing this parameter.

3. The *number of peaks* controlled the number of objectives in the search environment. Mathematically, this parameter controlled the number of layers of multivariate normal with single peaks.

4. The *distance between peaks*, in the range of [0,1], which prevented overlap of peaks when the function was generated with more than 1 peak.

For our experiments we wanted a single objective function, thus, number of peaks and distance between peaks are not varied and instead we focus on varying the ruggedness amplitude and smoothness between $[0.2, 0.8]$ with increments of 0.2 to limit observing functions of varying complexity. Apart from being able to control the complexity of the generated surface, our generator can produce multiple random surfaces of similar complexity, each setting of the function generator can be seeded to observe surfaces of similar complexity. These surfaces that have the same generator parameters, but only vary in their seed form a family of wildcatwells family. Using surfaces from the same family provide a useful tool for benchmarking optimization algorithms. A grid showing the variability in ruggedness as the smoothness and ruggedness amplitude parameters are changed is visualized in Fig. 3.1.

Algorithm 1 Constructing the Wildcat Wells search environment with given ruggedness (noise) amplitude (Rug_{amp}), smoothness (Smt), number of peaks (N) and distance between peaks (Rug_{freq}).

for $Rug_{amp}, Smt, N, Rug_{freq}$ **do**
    Get $X_{centers} = f(N, Rug_{freq})$
    Sample $\sum_i^N surf \sim \mathbb{N}(X_i)$
    Sample $Noise \sim$ OpenSimplex(Smt)
end for
Return $surf + noise \times g(Rug_{amp})$.

Next, we will provide a basic introduction to Bayesian optimization and provide specifics

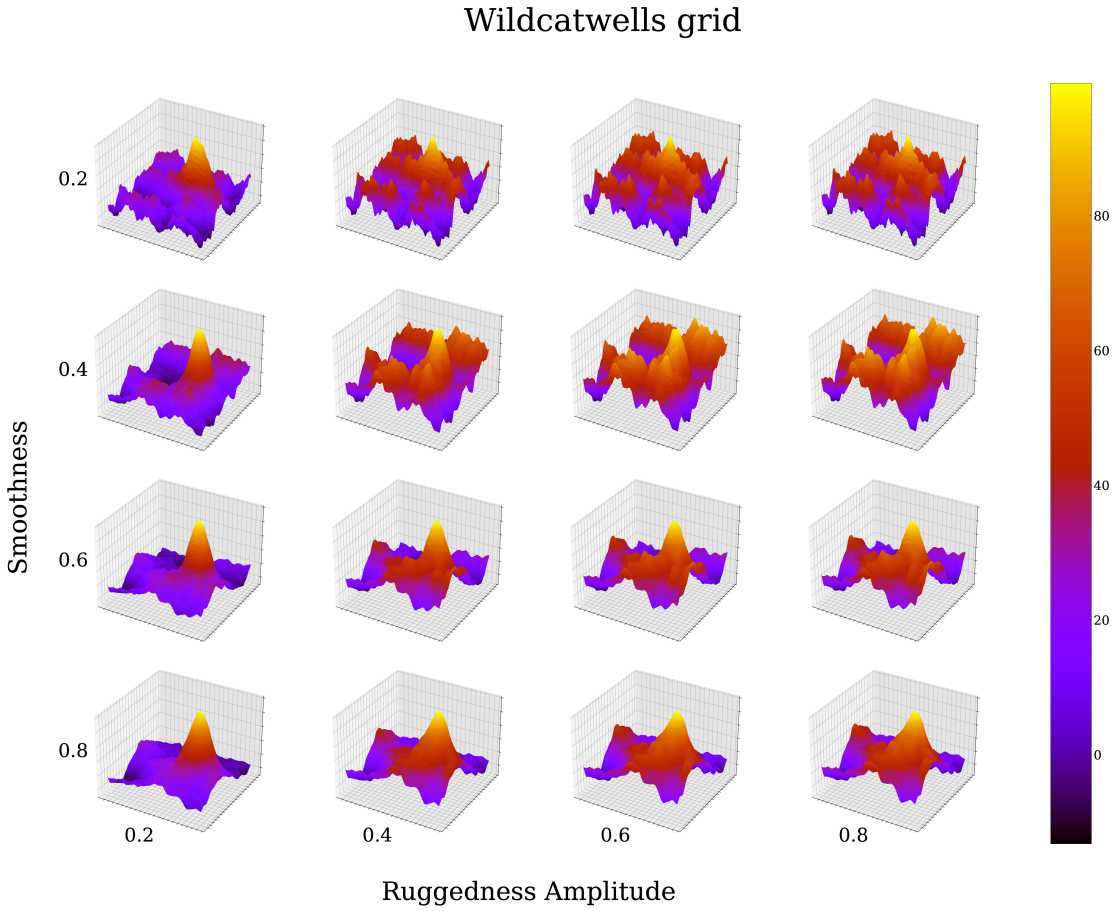


Figure 3.1: The grid plot shows how the landscape of wildcat wells changes as smoothness is varied between [0.2,0.8] in increments of 0.2 across the y-axis and ruggedness amplitude varied between [0.2,0.8] on the x-axis. So, the bottom-left plot in the grid corresponds to the smoothest family of wildcatwells functions and the top-right corresponds to the most rugged family.

on how BO is used in our experiments.

3.2 Bayesian Optimization

Bayesian optimization (BO) has emerged as a popular sample-efficient approach for optimization of these expensive black-box (BB) functions. BO models the black-box function using a surrogate model, typically a Gaussian process (GP) as seen in Eq. 3.1. The next design to evaluate is then selected according to an acquisition function. The acquisition function uses the GP

posterior and makes the next recommendation for function evaluation by balancing between exploration and exploitation. It allows exploration of regions with high uncertainty in the objective function, and exploitation of regions where the mean of the objective function is optimum. At each iteration, the GP gets updated according to the selected sample, and this process continues iteratively according to the available budget.

$$f(x) \sim \mathcal{GP}\left(\mu(x), k(x, x')\right), \tag{3.1}$$

,where $\mu(.)$ and $k(.,.)$ are the mean function and a real-valued kernel function encoding the prior belief on the correlation among the samples in the design space. In Gaussian process regression, the kernel function dictates the structure of the surrogate model we can fit. An important kernel for Bayesian optimization is the Matérn kernel, which incorporates a smoothness parameter ν to permit greater flexibility in modeling functions:

$$k_{\text{Matern}}(x, x') = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|x - x'\|}{\theta} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\|x - x'\|}{\theta} \right), \tag{3.2}$$

where σ_f^2 is the variance of the function, ν controls the smoothness of the function, θ is the length-scale parameter, and K_ν is the modified Bessel function.

Each data point in the context of Bayesian optimization is extremely expensive; thus, there is a need for selection of an informative set of initial samples for the optimization process. The core BO algorithm is relatively straightforward and involves iteratively selecting the next design to evaluate based on the current GP model. The algorithm for a simple BO model using a GP with Matern kernel is shown in Algorithm 2.

Algorithm 2 Bayesian Optimization with a Matern kernel and Expected Improvement

1: **Input:** objective function f, initial design set D, acquisition function A, GP model with Matern kernel
2: Initialize $X = D$, $y = f(X)$, and $m(\cdot)$ and $k(\cdot, \cdot)$ for the GP model
3: **while** not converged **do**
4: Update GP model parameters θ using marginal likelihood maximization
5: Compute acquisition function $A(X, \theta, y)$
6: Select next design $x_{next} = \arg\max_{x \in X} A(x, \theta, y)$
7: Evaluate objective function $y_{next} = f(x_{next})$
8: Update $X = X \cup x_{next}$, $y = y \cup y_{next}$, and GP model
9: **end while**

As seen above in Step 5, one needs to select an acquisition function. The acquisition function we used in this thesis for BO is the Expected Improvement (EI) function, which balances the trade-off between exploration and exploitation. The EI function is defined as:

$$EI(x) = \begin{cases} (y_{min} - \mu(x))\Phi(Z) + \sigma(x)\phi(Z) & \sigma(x) > 0 \\ 0 & \sigma(x) = 0, \end{cases} \tag{3.3}$$

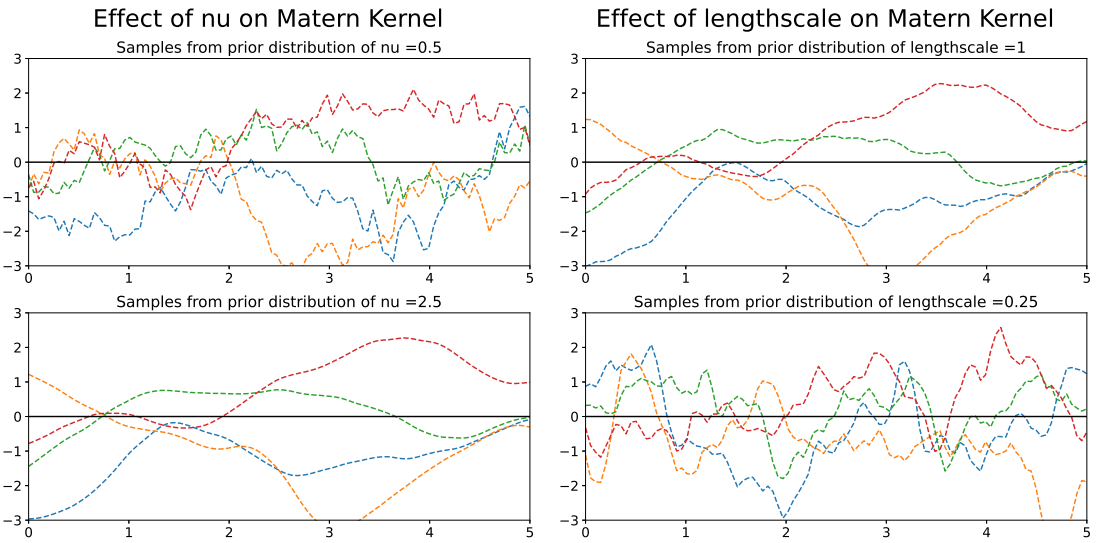
where $\mu(x)$ and $\sigma(x)$ are the mean and standard deviation of the GP at point x, y_{min} is the minimum observed value so far, and $Z = (y_{min} - \mu(x))/\sigma(x)$. $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal cumulative distribution function and density function, respectively. The intuition behind the EI function is to select points with high uncertainty (large $\sigma(x)$) and high potential for improvement (large $y_{min} - \mu(x)$).

We use the BOtorch [58] library to implement BO with EI acquisition function. BOtorch is an open-source Python library that provides a simple and modular framework for Bayesian optimization.

The *Model Building* advantage that we referred to earlier corresponds to learning these hyper-parameters of the Matérn kernel, which plays an important role in understanding the key

results of this thesis. So, let's take a closer look at these hyperparameters that Bayesian optimizers learn during the optimization process:

Lengthscale of the Matérn Kernel In Eq. 3.2, where θ is the lengthscale parameter of the kernel. This parameter controls the ruggedness expected by the Bayesian optimizer in the black box function being studied. The effects of lengthscale on GP can be seen in Fig. 3.2(b).



(a) Effect of changing the ν hyper-parameter on the Gaussian Process.

(b) Effect of changing the lengthscale hyper-parameter of the Matern Kernel. The figure has 2 similar GPs with shorter (bottom) and longer (top) lengthscales.

Figure 3.2: Effect of ν and lengthscale on Gaussian Process.

Output scale of Scale Kernel Output scale is used to control how the Matérn kernel is scaled for each batch. Since our Bayesian optimizer uses a single task GP, we do not use batch optimization. Thus, this parameter is unique for us and the way it's implemented using BoTorch can be seen Equation 3.4.

$$K_{scaled} = \theta_{scale} K_{orig} \tag{3.4}$$

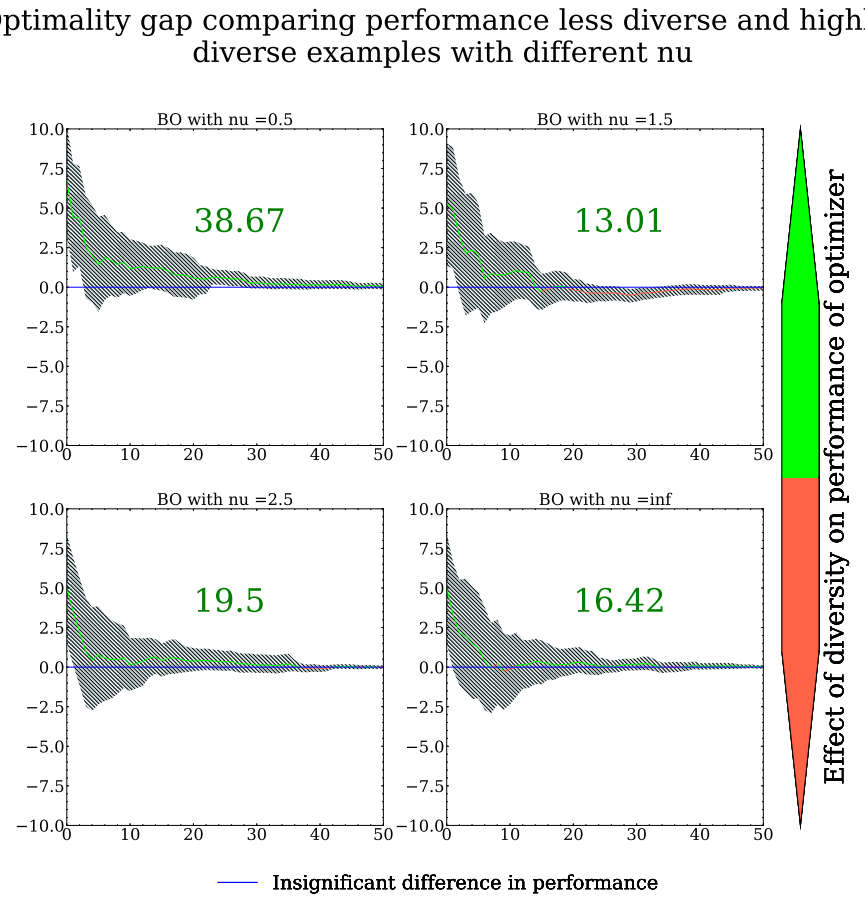


Figure 3.3: Grid plot showing how changing ν affects the relative performance of diverse and non-diverse initialization on Bayesian optimizers. To understand the plot better quantitatively, each subplot also has the Net Cumulative Optimality Gap (NCOG) for each value of ν. No trends are seen when relative performance of the diverse and non-diverse samples.

Noise for likelihood calculations The noise parameter is used to model measurement error or noise in the data. So, as the Gaussian Process gets more data the noise term decreases. So, ideally, this term should converge to 0 when the Bayesian optimizer has found an optimal value since our test functions did not have any added noise.

Constant for Mean Module This constant is used as the mean for the Normal distribution that forms the prior of the Gaussian Process as shown in Equation 3.1.

ν of the Matérn Kernel The hyper-parameter (ν) dictates how smooth or differentiable the function is. Changes in this parameter then influence the expectation of the Gaussian Process in terms of its acquisition function. A more differentiable function or a higher ν means that the acquisition function samples assuming a smoother Gaussian Process function. It can be seen in Fig. 3.2(a) how changes in ν changes the prior of the GP.

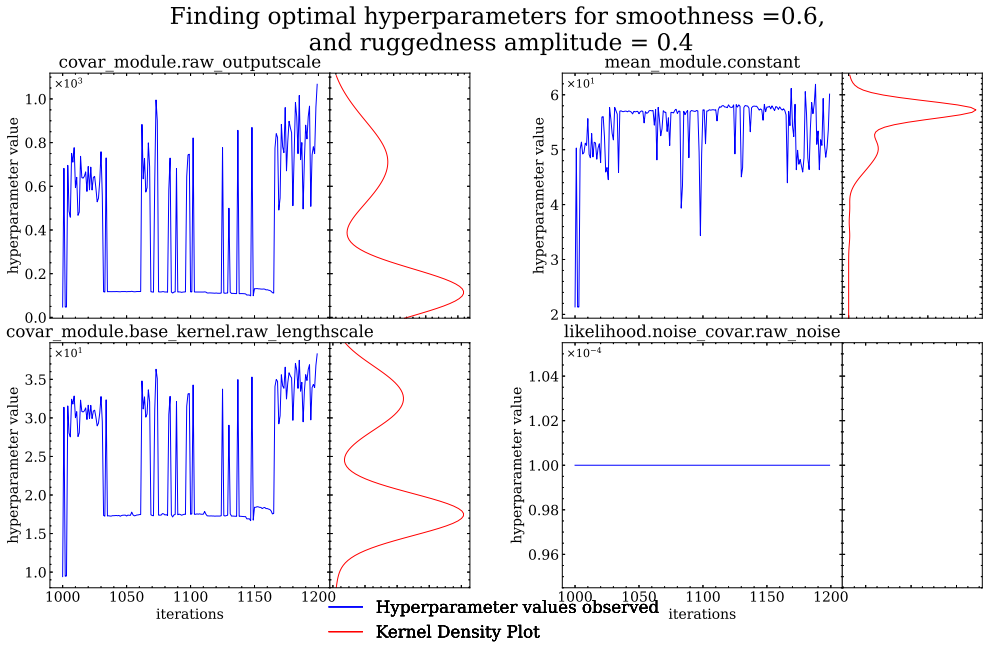


Figure 3.4: Figure depicts the first step in finding optimal hyperparameters for a wildcat wells function with smoothness 0.6 and ruggedness amplitude 0.4 and seed 88. Each hyperparameter in the grid plot has subsequent two adjacent plots. The observed hyperparameter values, when BOTorch is used to maximize the Marginal Log Likelihood, given 1000 to 1200 random points (left), the kernel density function derived from this data (right).

While ν controls the prior, μ and 'lengthscale' control how the data is scaled and thus indirectly controls the expectations of the GP. The effects of 'lengthscale' are similar to that of the parameter ν. Thus, we can conclude that 'lengthscale' can be used to control the expectations of the GP. Since, ν is not a parameter that is learned during the optimization process it does not have significant effect on "*Model Building advantage*". This can be seen in Fig. 3.3, even as ν is changed there is no significant change in the performance of the optimizer, and thus we can

conclude that ν is an insignificant factor in studying "*Model Building advantage*".

To provide some empirical evidence to the importance of 'lengthscale' as a hyper-parameter we will look at the performance of all hyper-parameters across diverse and non-diverse samples. But before we do that we have to understand what optimal hyper-parameter for a family of wildcat wells function is and how do we find them?

3.3 Finding the optimal BO kernel hyper-parameters for a given objective function

To compute the 'optimal hyper-parameter' we first use a Binary search method to discern a robust range (of 200 points) over which all families of wildcat wells functions has a noise parameter value of $< 10^{-5}$. This essentially means that Bayesian optimizer has found an optimal set of hyper-parameters for the Gaussian Process that accurately imitates the given black-box function.

This robust range for all the families of wildcat wells function used in the experiment was determined as 1000-1200 points.

Once, this range is determined the data is collected over the 200 points by maximizing the Marginal Log Likelihood for the Single Task GP model using BOTorch's '*fit-gpytorch-model*' [58] method. The resulting data (left side of every subplot) is the hyperparameters that BoTorch learns using the given data points (1000-1200). The resulting data (learnt hyperparameters) is then used to build a kernel density function as indicated by the red line-plot (right side of every subplot) next to the data observed over the 200 points in Fig. 3.4. Then using '*scipy.signal.find-peaks*' [59], peaks are found in the density function labeled by red dots in

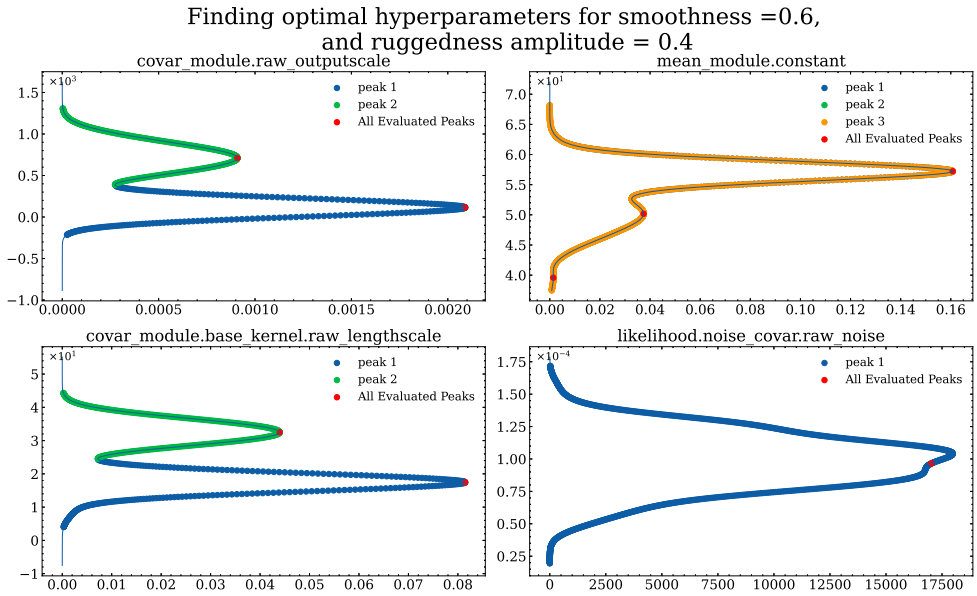


Figure 3.5: Figure depicting the second step in determining optimal hyperparameters. Each subplot shows the kernel density function learnt in the previous step. Each peak corresponds to the a potential optimal hyperparameter. The area under the curve of each peak is used as a tie-breaker. The shaded points that are used to calculate the area under the corresponding peak.

Fig. 3.5. Sometimes more than one peak is observed this is because there are multiple modes of hyperparameters that provide a stable solution for the problem. For the purpose of this thesis we only focus on extracting the most observed mode as our optimal hyperparameter.

To find the most observed mode, we use the width of the peaks in the kernel density function. The width of the peak is estimated by calculating a numerical gradient on the density function as seen in Fig. 3.5. The width of the each peak can be seen highlighted/labeled in each subplot using a different color. The peak with the largest area is selected as the optimal hyperparameter for the particular instance of wildcat wells function.

Following this methodology Fig. 3.6 shows how well do the initial samples learn the hyperparameters (less-diverse in blue and diverse in orange) vs the ground truth (blue horizontal lines). To the right of each box-plot in Fig. 3.6 is also 100 kernel density functions that have been used to estimate the 'optimal hyper-parameter' for each instance of that family (smoothness =0.6,

ruggedness amplitude =0.4) of wildcat wells function.

Now, as it can be seen in Fig. 3.6 the optimal noise hyper-parameter is close to 0 for all the instances in the family. While in the box-plot, the ones estimated using a sample size (k) of 10 are not. The performance for both diverse and non-diverse is relatively similar for this hyperparameter. This can be seen as the case for both the 'Mean function' (μ) and the 'Outputscale' as well. In contrast, 'lengthscale' is the only hyper-parameter that has varying performance across diverse and non-diverse samples. This is why in the next chapter to study 'modeling advantage' we will exclusively use the 'lengthscale' hyperaparameter. This same procedure will be used again to further generalize the result to other function settings in subsequent chapters.

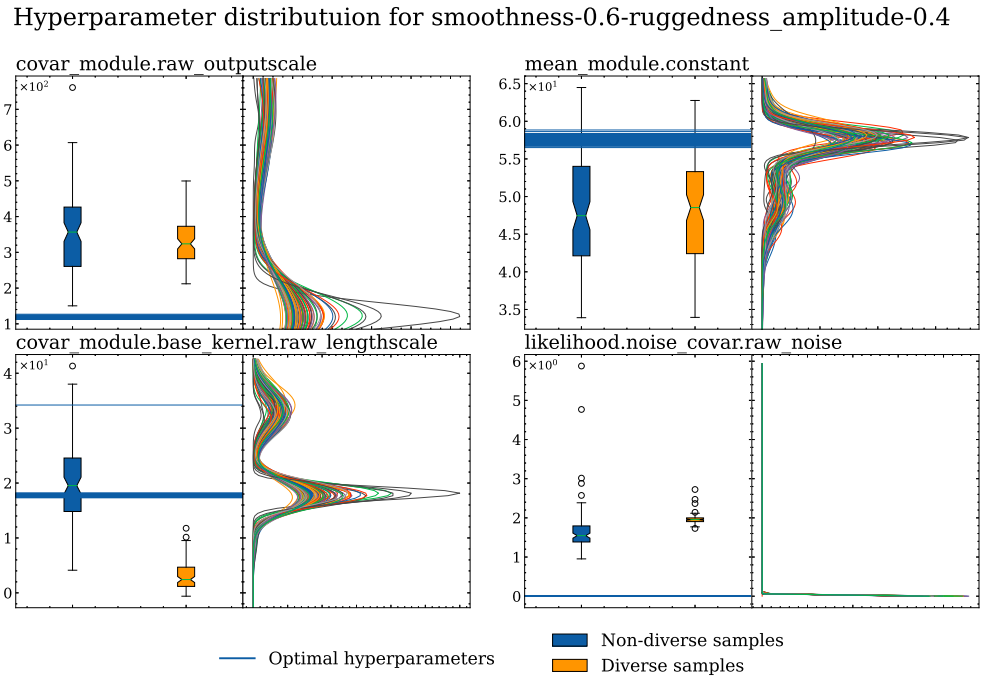


Figure 3.6: Box plots showing the distribution of different hyper-parameters of the Gaussian Process as learned by Bayesian optimizer when fitted with just the initial examples as training data. The shown hyper-parameters are specific to Wildcatwells configuration with smoothness = 0.6 and ruggedness amplitude =0.4. The data is collected over 100 seeds. The horizontal lines across the boxplot indicate the optimal hyper-parameters learned over 100 different seeds.

3.4 Diverse Sampling Method

To measure how the diversity of the initial set of points impacts the optimizer, we need to specify how we are assembling diverse or non-diverse sets. Our approach measures diversity of a set of examples using Determinantal Point Processes (DPP), which get their name from the fact that the probability of sampling a set from a DPP distribution is directly correlated to the determinant of a matrix referred to as an *L-ensemble* (as seen in Eq. 3.5) that correlates with the volume spanned by a collection or set of examples (Y) taken from all possible sets ($\mathcal{Y}$) given a diversity/similarity (feature) metric.

$$P(\mathbb{L}_Y) \propto det(K(\mathbb{L}_Y)) \tag{3.5}$$

Here $\mathbb{L}$ is the ensemble defined by any positive semi-definite matrix [27], and K is the kernel matrix. For sampling diverse examples, this positive semi-definite matrix is defined using similarity measures on pairs of examples. For this thesis, we use a standard and commonly used similarity measure defined using a Radial Basis Function (RBF) kernel matrix [60]. Specifically, each entry in $\mathbb{L}_Y$ for two examples with index i and j is:

$$[\mathbb{L}_Y]_{i,j} = \exp\left(-\gamma \cdot ||\mathbf{x}_i - \mathbf{x}_j||^2\right) \tag{3.6}$$

The hyper-parameter γ in the DPP kernel can be set in the interval $(0, \infty)$ and will turn out to be quite important in how well we can measure diversity. The next section explores this choice in more depth, but to provide some initial intuition: set γ too high and any selection of examples looks equally diverse compared to any other set, essentially destroying the discriminative power

of the DPP, while setting γ too low causes the determinant of $\mathbb{L}$ to collapse to zero for any set of cardinality greater than the feature-length of $\mathbf{x}$.

With $\mathbb{L}$ in hand, we can now turn Eq. 3.5 into an equality by using the fact that $\sum_{Y \subset \mathcal{Y}} det(\mathbb{L}_Y) = det(\mathbb{L} + I)$ [27], where I is an identity matrix of the same shape as the ensemble matrix $\mathbb{L}$. Then, using Theorem 2.2 from [27], we can write the $P(Y \in \mathcal{Y})$ as follows:

$$P(Y) = \frac{det(\mathbb{L}_{\mathbb{Y}})}{det(\mathbb{L} + I)} \tag{3.7}$$

This is the probability that a given set of examples (Y) is highly diverse compared to other possible sets ($\mathcal{Y}$)—that is, the higher $P(Y)$ the more diverse the set. The popularity of DPP-type measures is due to their ability to efficiently sample diverse examples of fixed size k. Sampling a set with k examples from a DPP is done using a conditional DPP called k-DPP [28]. k-DPP are able to compute marginal and conditional probabilities with polynomial complexity, in turn allowing sampling from the DPP in polynomial complexity. k-DPPs are also well researched and there exists several different methods to speed up the sampling process using a k-DPP [61, 62].

Our approach allows sampling in constant complexity however there is a trade-off in complexity in generating the DPP distribution, this can be seen in Fig. 3.8. The complexity for generating traditional DPP distributions is independent of 'k', while our approach has linear dependence on 'k'. Since, existing k-DPP approaches lack the ability to efficiently sample from different percentiles of diversity, which our approach permits without any additional cost.

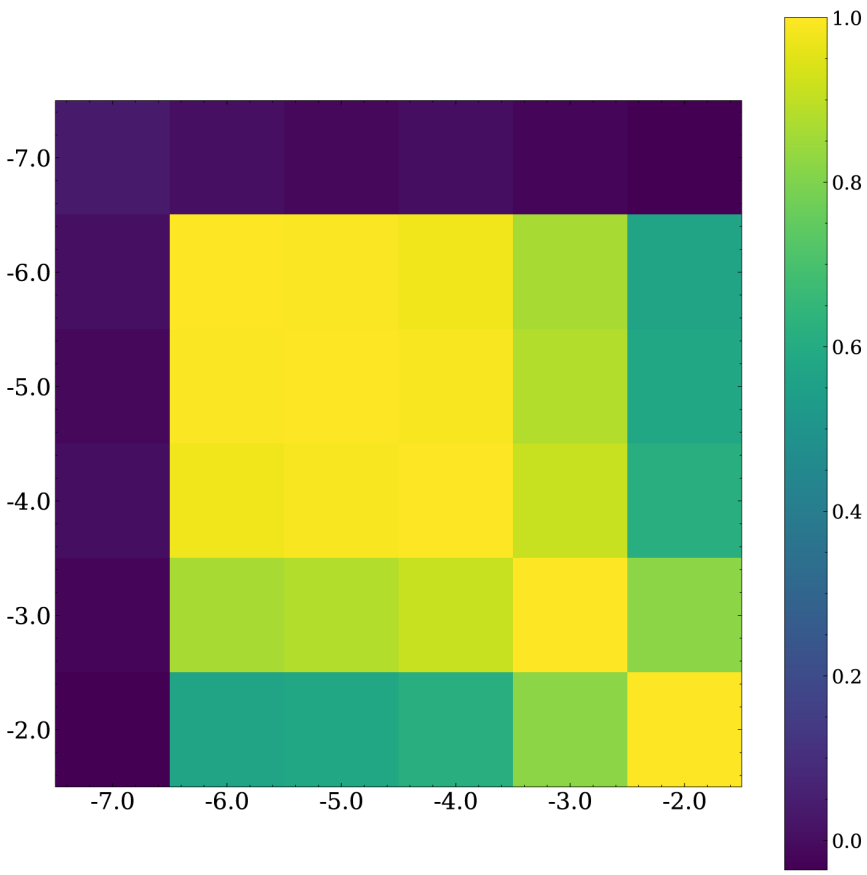


Figure 3.7: Correlation matrix showing the relative correlation between two gammas by comparing the way our DPP approach ranks 10,000 sampled sets of cardinality k=10. The gamma values in both axes here are logarithmic values with base 10.

3.4.1 Selecting the hyper-parameter for the DPP kernel

As mentioned above, the choice of γ impacts the accuracy of the DPP score, and when we initially fixed γ to $\frac{|Y_i|}{10}$, where Y_i is the set of data points over which the RBF kernel is calculating the DPP score as suggested by [63], the DPP seemed to be producing largely random scores. To select an appropriate γ we designed a kernel-independent diagnostic method for assessing the DPP kernel with four steps.

First, we randomly generate M samples of size k sets (think of these as random k-sized samples from $\mathcal{Y}$). Second, we compute their DPP scores for different possible γ values and then sort those M sets by that score. Third, we compute the rank correlation among these sets for different pairs of γ—intuitively, if the rank correlation is high (toward 1) then either choice of γ would produce the same rank orders of which points were considered diverse, meaning the (relative) DPP scores are insensitive to γ. In contrast, if the rank correlation is 0, then the two γ values produce essentially random orderings. This rank correlation between two different γ settings is the color/value shown in each cell of the matrix in Fig. 3.7. Large ranges of γ with high-rank correlation mean that the rankings of DPP scores are stable or robust to small perturbations in γ. Lastly, we use this "robust γ" region by choosing the largest range of γ values that have a relative correlation index of 0.95 or higher. We compute the mean of this range and use that as our selected γ in our later experiments. We should note that the functional range of γ is dependent on sample size (k), and so this "robust γ" needs to be recomputed for different initialization sizes.

The detailed settings for the results as seen in Figure 3.7 are as follows: the $M = 10000$; $k = 10$; $\gamma \in [e-7, e-2]$. The correlation matrix shows a range of γ with strongly correlating relative ordering of the test sets. All γ within this range provide a consistent ranking.

3.4.2 Our Sampling Approach

Our approach is designed to sample efficiently from different percentiles, this is made possible by creating an absolute diversity score. This score is generated by taking a $logdeterminant$ of the kernel matrix defined over the set Y. Randomly sampling the k-DPP space allows us to

bound errors in generating this absolute score through the use of concentration inequalities.

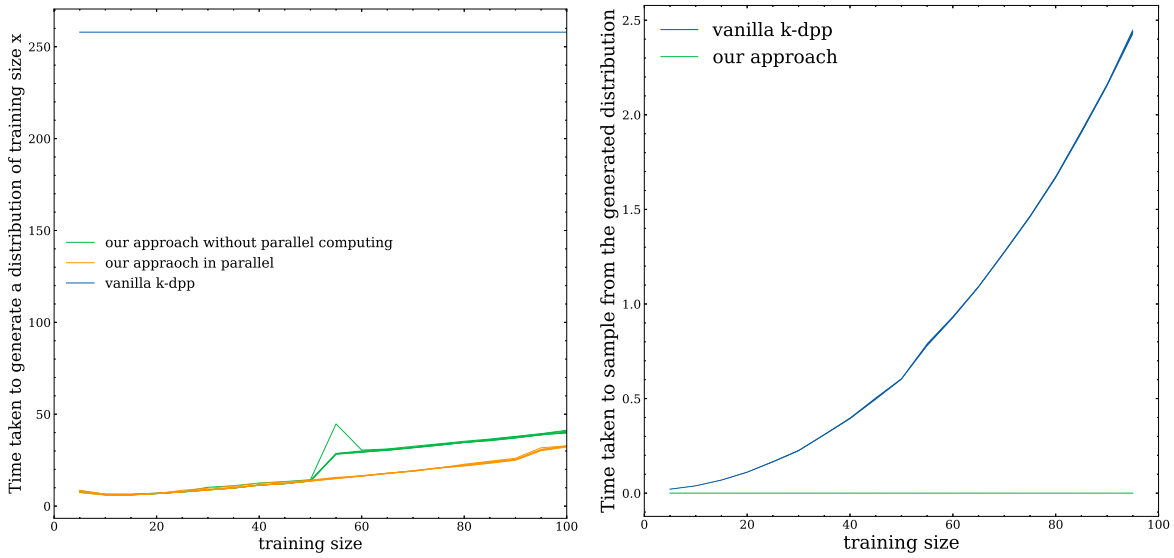
Our idea seeks to reduce the complexity of the sampling method and the construction time for DPP as well as investigate a Diverse sampling method that can generate both low-diversity and high-diversity examples. To do this we build on the work from [47] to rank and compare the diversity of the two sets. To define our diversity measure, let's assume $X \subset \mathbb{R}^{\mathcal{F}}$, where $|\mathcal{F}|$ is the number of features of X. Then we can define a set of examples as $S_Y^k \subset X$ of size k. This means $S_{Y_i}^k \in \mathbb{R}^{\mathcal{F}} \times \mathbb{R}^k$, then using a similarity measure (RBF kernel) W on this set, we can define the DPP score for a set $S_{Y_i}^k$ as follows:

$$f(W_{Y_i}) = \frac{\log(det(K(W_{Y_i}))) - \left(\sum_i^{\|S^k\|} log(det(K(W_{Y_i})))\right)}{\sqrt{\sum_i^{\|S^k\|}(\log(det(K(W_{Y_i}))) - \left(\sum_i^{\|S^k\|} log(det(K(W_{Y_i})))\right))^2}} \tag{3.8}$$

As we can see in Eq. 3.8, where $\|S^k\| = \binom{\left(\Pi_i^{dim(X)} dim(\mathcal{F}_i)\right)}{k}$ the number of sets or cardinality of the distribution $\|S^k\|$ needed to be sampled grows combinatorially with the changes in the size of the example space for Xs, and the size of the set k. For example, for a $X \in \mathbb{Z}^2$ where each feature $\mathbb{Z}_i \in [0, 100]$ the number of possible sets of size k is given by $\binom{100\times100}{k}$, thus normalizing the distribution using a mean and standard distribution is an expensive task. We can re-write Eq. 3.8 in words as follows:

$$\text{DPP Score}(S_{Y_i}^k) = \frac{DPPScore(K(W_{Y_i}) - \text{mean score}}{\text{s.d. of DPP scores for the k-HDPP}}$$

Given the above, the sampling method for our DPP approach is straightforward. Based on the constructed sampling distribution, our approach samples randomly from either above a



(a) Compares the construction time for regular k-DPP distribution with different size of samples

(b) the sampling time time for regular k-DPP vs our approach as the size of k is increased from 5 to 100.

Figure 3.8: Compares the relative performance/speed-up of our method over the traditional k-dpp methods. The figure contains two plots showing the tradeoff between the two methods. In the traditional method constructing the DPP distribution is costly but generating a distribution is only dependent on the number of points in X, and independent of training size (k). While, sampling from a k-DPP has a polynomial complexity on the training size (k), while both these facts are inverted for our approach.

certain percentile or below a certain percentile. As shown in Fig. 3.8(b), our approach's sampling time is faster than that of a regular k-DPP, where the cost of sampling increases as a function of training size (k). Conversely, generating the distribution for our approach is dependent on 'k' this is because we are sampling same number of sets but it has now more elements, while the for k-DPP(s) the distribution generated is over the whole data and computes correlations in all data, thus sampling the k most diverse points doesn't require re-generating the distribution. Our approach's biggest benefit is the ability to draw examples of different levels of diversity. Using our approach this is as simple as sampling from different percentiles of the distribution.

A clear shortcoming of this approach is the need to generate the distribution whenever the k is changed. But, because of the faster construction speed for our approach, this cost outweighs using a k-DPP. Another, shortcoming our approach faces is the limited number of examples that

Algorithm 3 Generating a ranked distribution of example sets with Determinantal Point Process (DPP) approach [27], M is batch size (10000). S^k is a combinatorial set defined on a finite set $X \in \mathbb{R}^2$, where each element $S^k_{Y_i} \in S^k$ is k elements long.

1: **for** $i \in range(M)$ **do**
2: Sample $S^k_{Y_i} \sim \mathbb{IIID}(S^k)$ [identically sampling unordered sets without replacement]
3: Calculate $g(S^k_{Y_i}) = g_{y_i}$ and append this to $Scores_{S^k}$
4: **end for**
5: Return DPP Score of sets of examples = $\frac{Score_{S^k} - mean(Score_{S^k})}{s.d.(score_{S^k})}$.

can be drawn from the distribution, which requires us to construct a new distribution if more than M examples need to be drawn.

The uniqueness of our approach lies in our ability to upper bound the error on the generated DPP scores, and thus our approach can provide certain guarantees on whether the sampled S^k_Y is in fact from the percentile that the method claims it is from.

3.4.3 Upper bounding DPP sampling errors

The guarantee is based on method's independence of choosing the S^K_Y from a combinatorially large set. For IID sampling each set, $S^K_{Y_i}$, needs to be sampled independent of the other and the sampling should be done with replacement. But since the distribution of S^k needs to mirror that of a k-DPP, all the sets in the space are sampled over X without replacement and are unordered because DPP scores for two S^k with the same points (Y) will always correspond to the same score. Thus, sampling IID on S^k means identically sampling unordered sets of X without replacement.

If we sample the sets $S^k_{Y_i}$ such that they are Independent Identical Distributed (I.I.D.) sets, then we can upper bound the Expected Value of population mean through the use of Hoeffding's

inequality: Eq. 3.9 as discussed in [64]. The inequality states that if a distribution is sampled using i.i.d random variables, we can then put a bound on the Error for estimating Expected Values of the population mean ($|\mathbb{M}_n = \frac{1}{n}\sum_i^n [M_i]|$), where $\mathbb{M}_n$ is the mean of the distribution of all sets in S^k of size n.

$$\mathbb{P}\{\|\mathbb{M}_n - \mathbb{E}(S^k)\| \leq \epsilon\} \geq 1 - 2 \cdot exp\left\{\frac{-2 \cdot n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\} \tag{3.9}$$

Using Eq. 3.9 we can guarantee the probability of this error to be some $1 - \delta$, where the δ term is given by the exponential. This allows us to limit the cardinality of the $|S^k|$ to M given we choose an ϵ. Based on this guarantee a schematic explanation for the construction of our sub-distribution using the approach detailed till now is then documented well in Algorithm 3.

With these tools in hand, the next chapter can now move to the core experiments that measure the effect of diverse initializations on BO performance.

Chapter 4: Does Diversity Affect Optimization Convergence?

To test the effects of diversity of initial samples on optimizer convergence, we first generated a set of initial training samples of size (k) 10 either from low (5^{th} percentile of diversity) or high diversity (95^{th} percentile of diversity) using our procedure in Sec. 2. Next, we created 100 different instances of the wildcat wells function with different randomly generated seeds for each cell in a 4x4 factor grid of 4 values each of the smoothness and ruggedness amplitude parameters of the wildcat wells function (ranging from 0.2 to 0.8, in steps of 0.2). For simplicity here, we refer to these combinations as families of the wildcat wells function. This resulted in 1600 function instances.

4.1 When Bayesian Optimization is allowed to fit its kernel hyperparameters

Our first experiment consisted of 200 runs of the Bayesian Optimizer within each of the smoothness-ruggedness function families, where each run consisted of 100 iterations, and half of the runs were initialized with a low-diversity training sample, and half were initialized with a high-diversity training sample. Importantly, as we will see, by default our BO library (BoTorch) automatically optimizes the kernels hyper-parameters, as we described in the last chapter.

We then compared the cumulative optimality gap across the iterations for the runs with low-diverse initializations and high-diverse initalizations within each smoothness-ruggedness combi-

nation family. We did this by computing bootstrapped mean and confidence intervals within each low-diverse and high-diverse sets of runs within each family. Given the full convergence data, we compute a Cumulative Optimality Gap (COG) which is just the area under the Optimality Gap curve for both the 5^{th} and 95^{th} diversity curves. Intuitively, a larger COG corresponds to a worse overall performance by the optimizer. Using these COG values we can numerically calculate the improvement of the optimizer in the 95^{th} percentile. The net improvement of COG value while comparing the 5^{th} and 95^{th} percentile is also presented as text in each subplot in Figure 4.1.

To confirm what we observed was not limited to 2 dimensions we decided to run our current study with wildcatwells in 3 dimensions. To make the results comparable in a single figure for both 2D and 3D case it was necessary to limit the variability of ruggedness from a 4x4 grid to 3 levels of 'ruggedness'. These 'levels of ruggedness' are 'low', 'medium' and 'high', which correspond to (smoothness : 0.8, ruggedness amplitude : 0.2), (smoothness : 0.4, ruggedness amplitude : 0.4) and (smoothness : 0.2, ruggedness amplitude : 0.8) respectively.

4.1.1 Results

As Figs. 4.1,4.2 show, the Cumulative Optimality Gap does not seem to have a consistent effect across the grid. Diversity produces a positive convergence effect for some cells, but is negative in others. Moreover, there are wide empirical confidence bounds on the mean effect overall, indicating that should an effect exist at all, it likely does not have a large effect size. Changing the function ruggedness or smoothness did not significantly modulate the overall effect. As expected, given sufficient samples (far right on the x-axis) both diverse and non-diverse initializations have the same optimality gap, since at that point the initial samples have been crowded out by the new

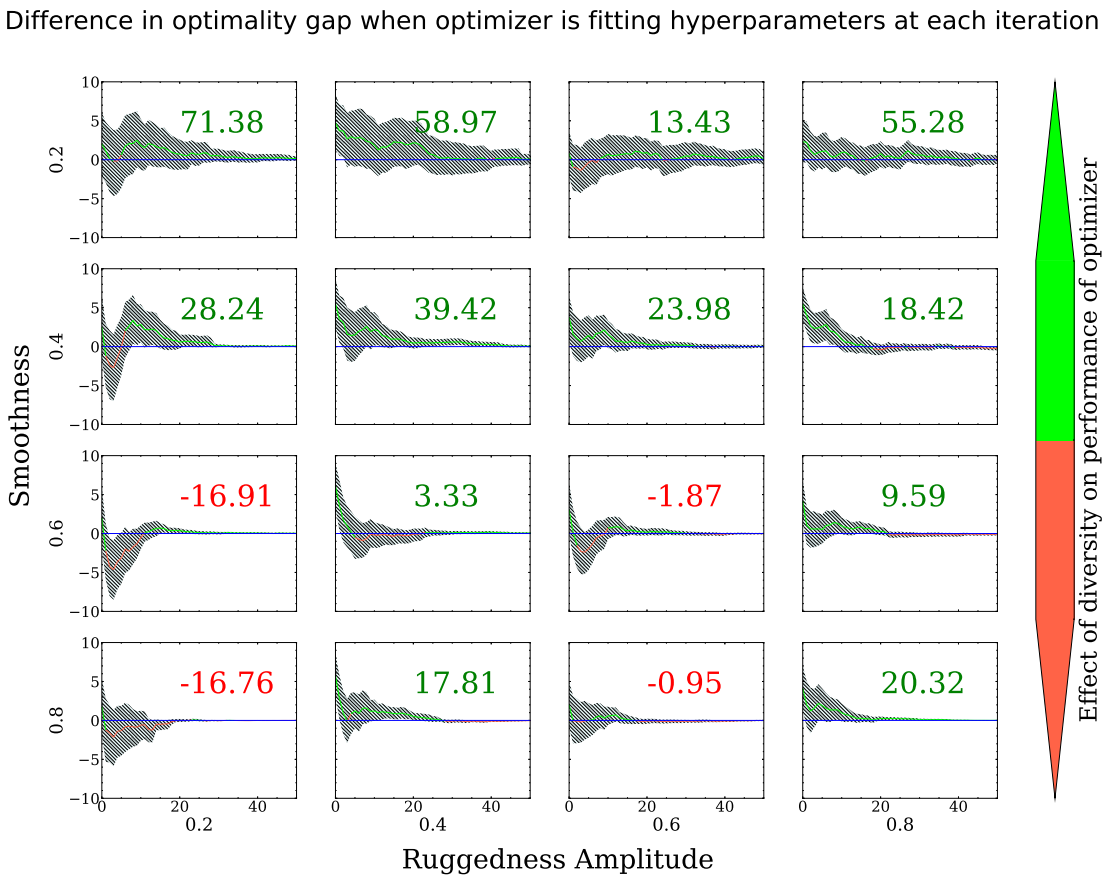


Figure 4.1: Experiment 1: Optimality gap grid plot showing the difference in current Optimality Gap between optimizers initialized with 5th vs 95th percentile diverse sample (y-axis) as a function of optimization iteration (x-axis). The different factors in the factor grid plot the effects of diversity as the noise amplitude and smoothness are varied in the range [0.2,0.8]. Each plot also has text indicating the Net Cumulative Optimality Gap (NCOG), a positive value corresponds to a better performance by high diversity samples compared to the low diversity samples. The plot shows that BO benefits from diversity in some cases but not others. There is no obvious trends in how the NCOG values change in the grid. The results are further discussed in §4

samples gathered by BO during its search.

4.1.2 Discussion

Overall, the results from Figs. 4.1,4.2 seem to indicate that diversity helps in some cases and hurts in others, and regardless has a limited impact one way or the other. This seems counter to the widespread practice of diversely sampling the initial input space using techniques like LHS. Figure 4.1 shows that it has little effect.

Why would this be? Given decades of research into initialization schemes for BO and

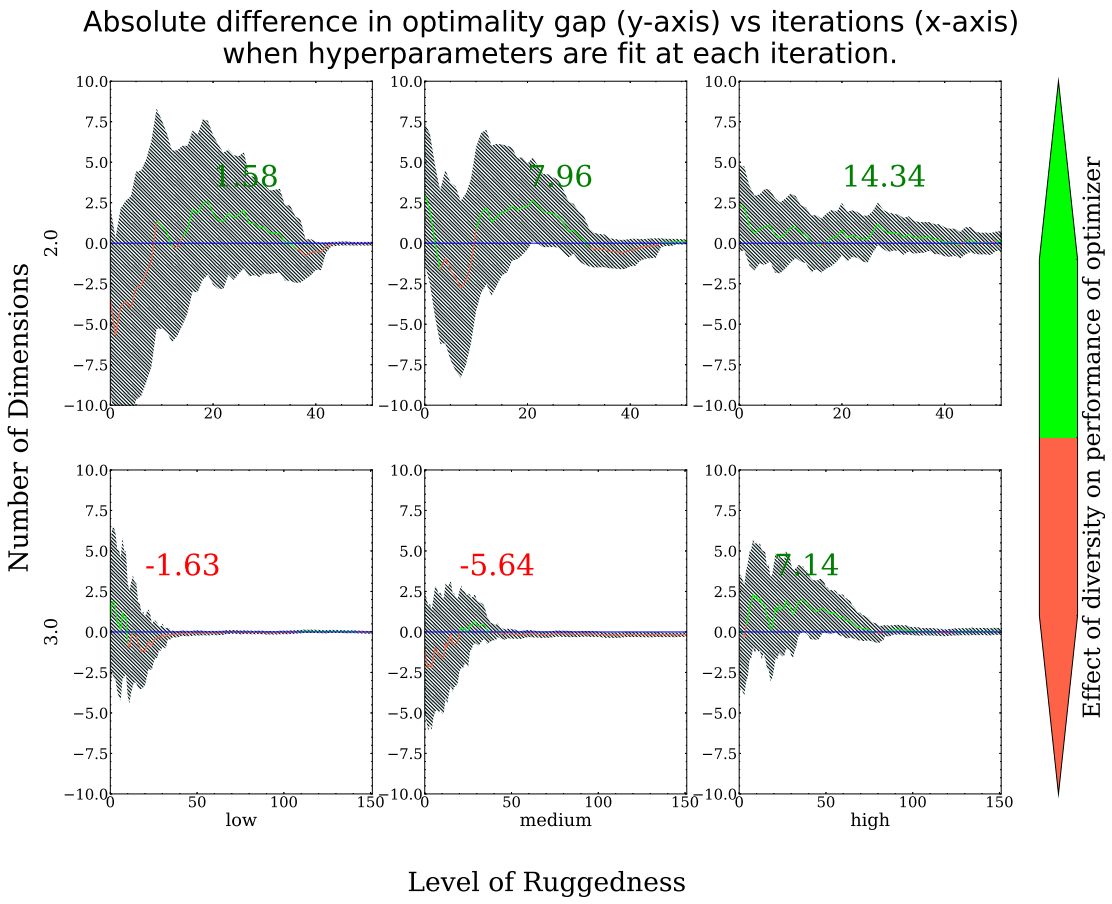


Figure 4.2: Optimality gap grid plot showing the difference in current Optimality Gap between optimizers initialized with 5^{th} vs 95^{th} percentile diverse sample (y-axis) as a function of optimization iteration (x-axis). The different factors in the factor grid plot are the dimensions across the rows and the ruggedness level across the columns. Each plot also has text indicating the Net Cumulative Optimality Gap (NCOG), a positive value corresponds to a better performance by high diversity samples compared to the low diversity samples. The plot shows that BO benefits from diversity in some cases but not others. There is no obvious trends in how the NCOG values change in the grid. The results are further discussed in Sec. 4

Optimal Experiment Design, we expected diversity to have at least some (perhaps small but at least consistent) positive effect on convergence rates, and not the mixed bag that we see in Figs. 4.1,4.2. How were the non-diverse samples gaining such an upper hand when the diverse samples had a head start on exploring the space—what we call a *Space Exploration* advantage?

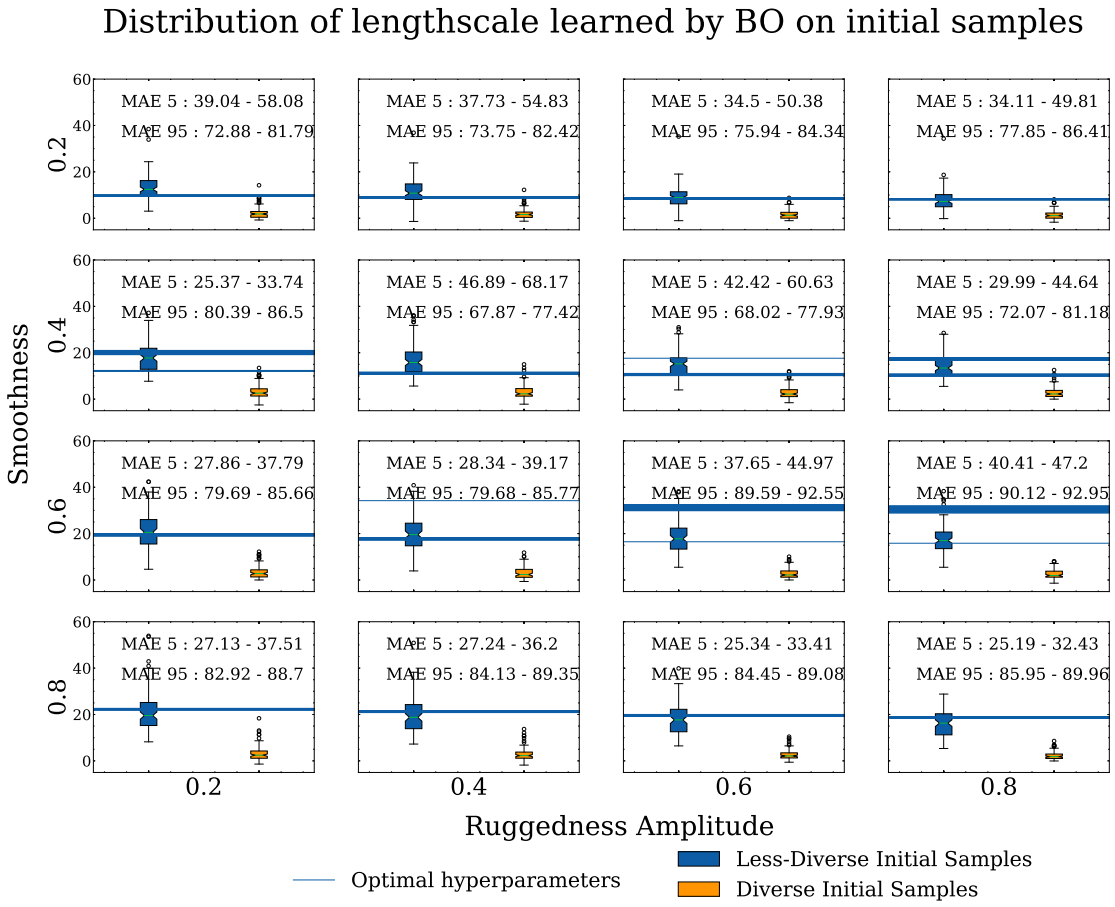


Figure 4.3: Experiment 2: Box plot showing the distribution of 'Lengthscale' hyper-parameter learned by BO when initiated with diverse (orange) and non-diverse samples (blue) for 16 different families of wildcat wells functions of the same parameters but 100 different seeds. The optimal hyper-parameter for each of the 100 wildcat wells instances from each family is also plotted as horizontal (blue) lines—in many but not all cases these overlap. Each cell in the plot also has the 95^{th} percentile confidence bound on Mean Absolute Error (MAE) for both diverse and non-diverse samples. The results show that MAE confidence bounds for non-diverse samples are smaller compared to diverse samples for all the families of wildcat wells function. Thus, indicating a presence of Model Building advantage for non-diverse initial samples. The results of this figure are further discussed in §4.2

4.2 Do Lower Diversity Samples Improve Hyper-parameter Posterior Convergence?

After reviewing the results from Fig. 4.1, we tried to determine why the Space Exploration advantage of diversity was not helping BO as we thought it should. We considered as a thought experiment the one instance where a poorly initialized BO model with the same acquisition function might outperform another: if one model's kernel hyper-parameter settings were so grossly

incorrect that the model would waste many samples exploring areas that it did not need to if it had the correct hyper-parameters.

Could this misstep be happening in the diversely sampled BO but not in the non-diverse case? If so, this might explain how non-diverse BO was able to keep pace: while diverse samples might give BO a head start, it might be unintentionally blindfolding BO to the true function posteriors, making it run ragged in proverbial directions that it need not. If this hypothesis was true, then we would see this reflected in the comparative accuracy of the kernel hyper-parameters learned by the diverse versus non-diverse BO samples. Our next experiment set out to test that hypothesis.

4.2.1 Methods

The key difference from our study above is that, rather than comparing the overall optimization convergence, we instead focus on how the initial samples' diversity affects BO's hyper-parameter posterior convergence, and compare how far each is from the "ground truth" optimal hyperparameters.

As with the above, we used the same smoothness and ruggedness amplitude families of the wildcat wells function. To then generate the data for each instance in one of these families, we sampled 20 sets of initial samples. Half of the sampled 20 sets were low (5^{th} percentile of diversity) and the other half from high diversity (95^{th} percentile of diversity) percentiles.

For each initial sample, we then maximized the GP's kernel Marginal Log Likelihood (via BOTorch's GP fit method). We then recorded the hyper-parameters obtained for all 20 initial samples. The mean of the 10 samples from low diversity was then used as one point in the

box plot's low diversity distribution as seen in Fig. 4.3. We then repeated this process for the high-diversity initial samples. Each point in the box plot can be then understood as the mean hyper-parameter learned by BOTorch given just the initial sample of size (k) 10 points. To get the full box plot distribution for each family the above process is repeated over 100 seeds and Fig. 4.3 provides the resulting box plot for both diverse and non-diverse initial samples for all the 16 families of wildcat wells function as described in Experiment 1 [1].

To provide a ground truth for the true hyper-parameter settings, we ran a Binary search to find the size of the sample ($k_{optimal}$) for which BO's kernel hyper-parameters converged for all families. The hyper-parameter found by providing $k_{optimal}$ amount of points for each instance in the family was then plotted as a horizontal line in each box plot. An interesting observation is that some families have non-overlapping horizontal lines. This is because for some families there are more than one modes of 'optimal hyper-parameters'. The mode chosen as the 'optimal hyper-parameter' is the more observed mode. The process for finding the 'optimal hyper-parameter' and which mode is chosen as the optimal hyper-parameter has been described later.

If an initial sample provides a good initial estimate of the kernel hyper-parameter posterior, then the box plot should align well or close to the horizontal lines of the true posterior. Figure 4.3 only shows the results for the Matérn Kernel's Lengthscale parameter, given its out-sized importance in controlling the GP function posteriors compared to the other hyper-parameters (*e.g.*, output scale, noise, *etc.*), which we do not plot here for space reasons. We provide further details and plots for all hyper-parameters in the Appendix.

To quantify the average distance between the learned and true hyper-parameters, we also

[1] At several places throughout this thesis, experiment and the word study has been used interchangeably. Experiment 1 refers to our study in Sec. 4, Experiment 2 refers to our study in Sec. 4.2 and Experiment 3 refers to the study in Sec. 4.4

plot on Fig. 4.3 the Mean Absolute Error (MAE) for both highly diverse (95^{th}) and less diverse (5^{th}) points. The MAE is the sum of the absolute distance of each predicted hyper-parameter from the optimal hyper-parameter for the particular surface of each wildcat wells function. The range as seen in each cell in Figure 4.3 corresponds to a 95^{th} percentile confidence bound on the Mean absolute error across all the 100 runs.

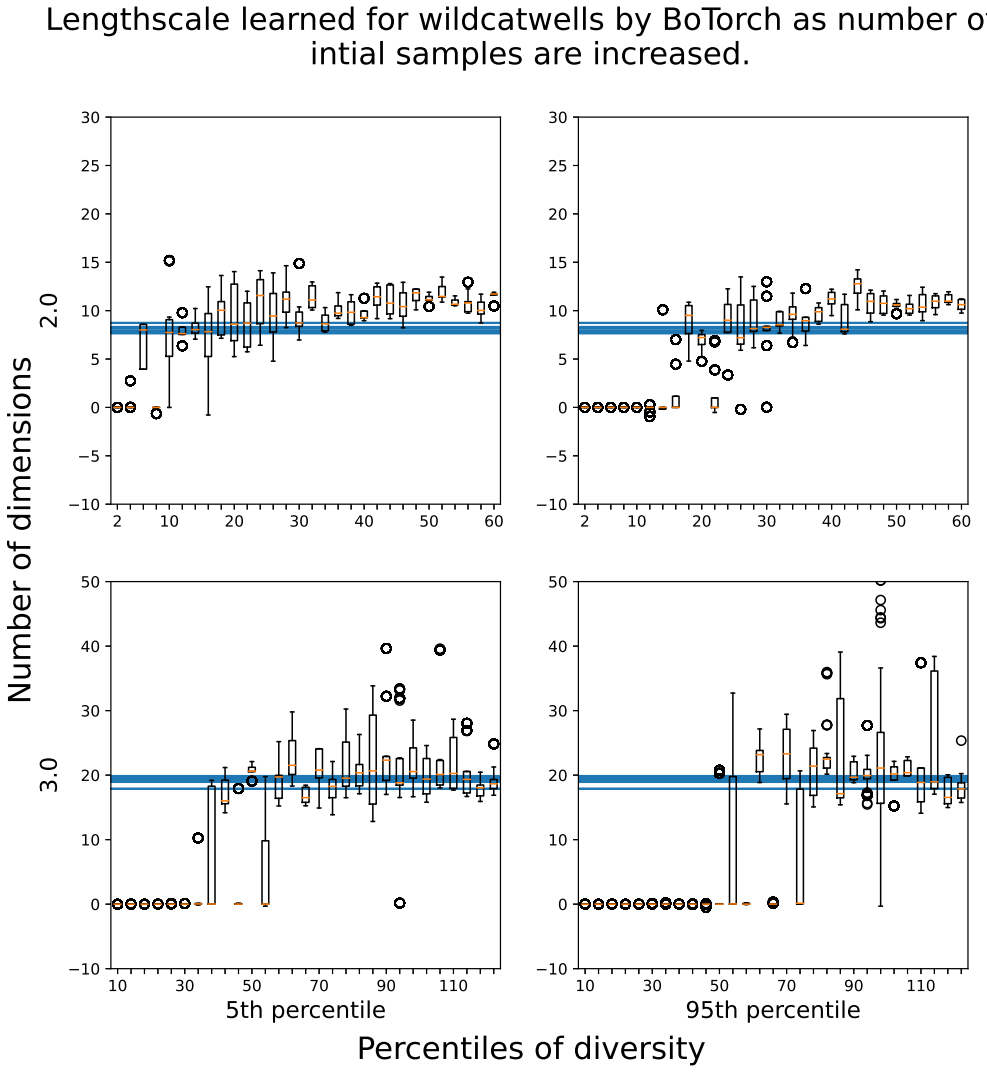


Figure 4.4: Box plot showing the lengthscale parameter as learned by wildcatwells with 'high' level of ruggedness in 2D and 3D as the training samples are increased. The plot also confirms the existence of a "modeling advantage" for training samples of a particular size. The results are further discussed in Sec. 4.3

4.2.2 Results and Discussion

The results in Figure 4.3 show that the MAE values for low diversity samples are always lower compared to the MAE for high diversity samples. This general behavior is also qualitatively visible in the box plot. This means that after only the initial samples, the non-diverse samples provided much more accurate estimates of the kernel hyper-parameters compared to diverse samples. Moreover, BO systematically *underestimates* the correct lengthscale with diverse samples—this corresponds to the diverse BO modeling function posteriors that have higher frequency components than the true function actually does.

This provides evidence for the *Model Building* advantage of non-diverse samples that we defined in Sec. 3.2. It also confirms our previous conjecture from the thought experiment that diverse samples might be impacting BO by causing slower or less accurate convergence to the right BO hyper-parameters. The Space Exploration advantage of the diverse samples helps it compensate somewhat for its poor hyper-parameters, but BO trained with non-diverse samples can leverage the better hyper-parameters to make more judicious choices about what points to select next.

We did not see major differences in the other three kernel hyper-parameters such as Output Scale, Noise, or the Mean Function; however, this is not surprising, since BO is not highly sensitive to any of these parameters and the lengthscale parameter dominates large changes in BO behavior.

Comparing the different smoothness and ruggedness settings, when the function is more complex (the top right of the grid at low smoothness and high ruggedness amplitude values) the function's lengthscale is lower and closer to the value learned by the diverse samples. Looking

at the low diversity MAE values ('MAE 5'), we can see they are much closer to those of the high diversity samples ('MAE 95'), in contrast to when the function is less complex (bottom left side of the grid). Under such conditions, low diversity samples lose some of the relative Model Building advantage they have over high diversity samples. This conjecture aligns with Experiment 1 (Fig 4.1) where the COG values on the top right part are positive while those on the bottom left are negative.

4.3 How is this phenomenon affected by the initial training set size?

While trying to replicate the results for the 3D case we observed that the 'modeling advantage' we observed for less diverse examples was also influenced by the number of examples in the initial set. This was because if we initialized the 3D case with the same number of initial samples as the 2D case, the optimizer in the 3D case would not be able to accurately estimate the appropriate hyperparameters regardless of the sampling (less diverse or diverse) method and would just set the hyperparameters to zero.

This is perhaps obvious if we think about how space coverage degrades for a fixed number of samples as we increase the dimensionality of a design space. What we observed, and show below in Fig. 4.4, is that there are essentially three "initial sample size regimes" that determine whether or not non-diverse sampling can use its 'modeling advantage', although this advantage exists in both the 2D and 3D case:

1. Sample-deficient: This is when we provide each optimizer with too few initial examples, such that irrespective of that set's diversity the BO will not be able to meaningfully learn hyperparameters and will instead set them to zero. For example, in Fig. 4.4 bottom, with

fewer than 26 initial samples, both the 5th and 95th percentile samples cannot provide good estimates of the kernel hyper-parameters

2. The 'modeling advantage' region: With this number of samples, the 5th percentile is able to reasonably estimate the hyperparameter values but the 95th struggles to do so. For example, in Fig. 4.4 top (2D), we can observe this at 10 samples, which, by coincidence, was the original setting for our 2D example in our initial manuscript. We see that in Fig. 4.4 bottom (3D) this transitions somewhere between 35 to 75 initial samples. In this region, 5th percentile sampling can exercise its modeling advantage while the 95th percentile still does not have enough initial samples to consistently and accurately estimate the kernel hyper-parameters.

3. Sample-saturated: In this region, the shear number of initial points we provide BO is sufficiently high such that it can estimate the kernel hyper-parameters well, regardless of whether the initial points are diverse or not. For example, in Fig. 4.4 top, this occurs after around 40 initial samples. In Fig. 4.4 bottom this occurs after around 100 initial samples. In this 'sample-saturated' case, the modeling advantage of non-diverse sampling disappears, often because this is a sufficient number of points that the optima become easy to find at that point (see Fig. 4.2 where the BO often converges at those same number of samples).

Once, the training size was fixed to 40 examples for the 3D case we can see that Figure 4.5 demonstrates our hypothesized Model Building advantage that non-diverse initial samples confer to BO. But how do we know that this is the actual causal factor that accelerates BO convergence, and not just correlated with some other effect? If correct, our conjecture posits a natural testable hypothesis: if we fix the values of the hyper-parameter posteriors to identical values between the

non-diverse and diverse samples and do not allow the BO to update or optimize them, then this should effectively eliminate the Model Building advantage, and diverse samples should always outperform non-diverse samples.

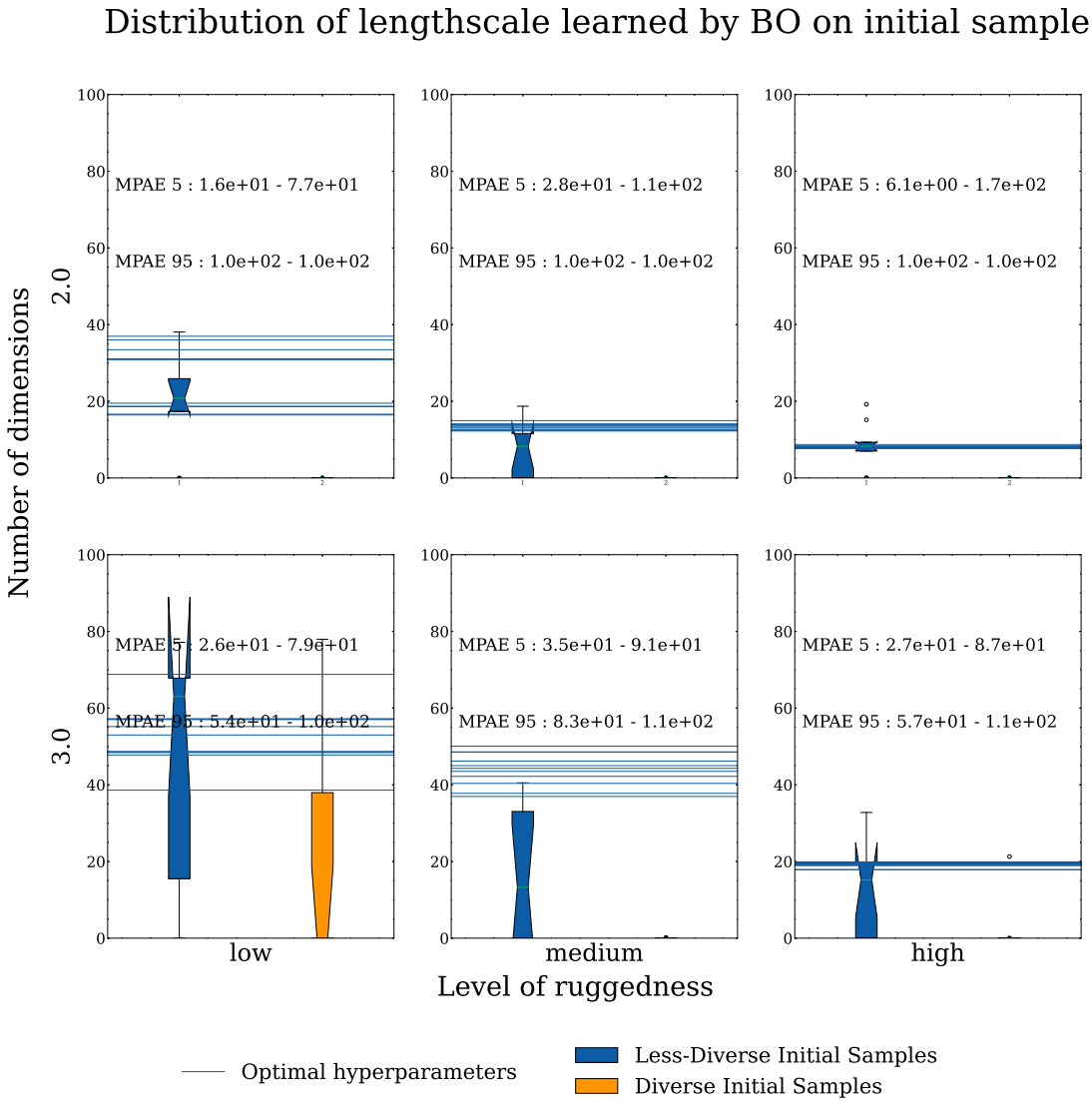


Figure 4.5: Box plot showing distribution of 'Lengthscale' hyper-parameter learned by BO when initiated with diverse (orange) and less-diverse samples (blue) for 3 different families of wildcat wells functions of the same parameters but 100 different seeds in each dimension. The optimal hyper-parameter for each of the 100 wildcat wells instances from each family is also plotted as horizontal (blue) lines—in many but not all cases these overlap. Each cell in the plot also has the 95^{th} percentile confidence bound on Mean Absolute Error (MAE) for both diverse and non-diverse samples. The results show that MAE confidence bounds for non-diverse samples are smaller compared to diverse samples for all the families of wildcat wells function. Thus, indicating a presence of Model Building advantage for non-diverse initial samples. The results of this figure are further discussed in Sec. 4.3

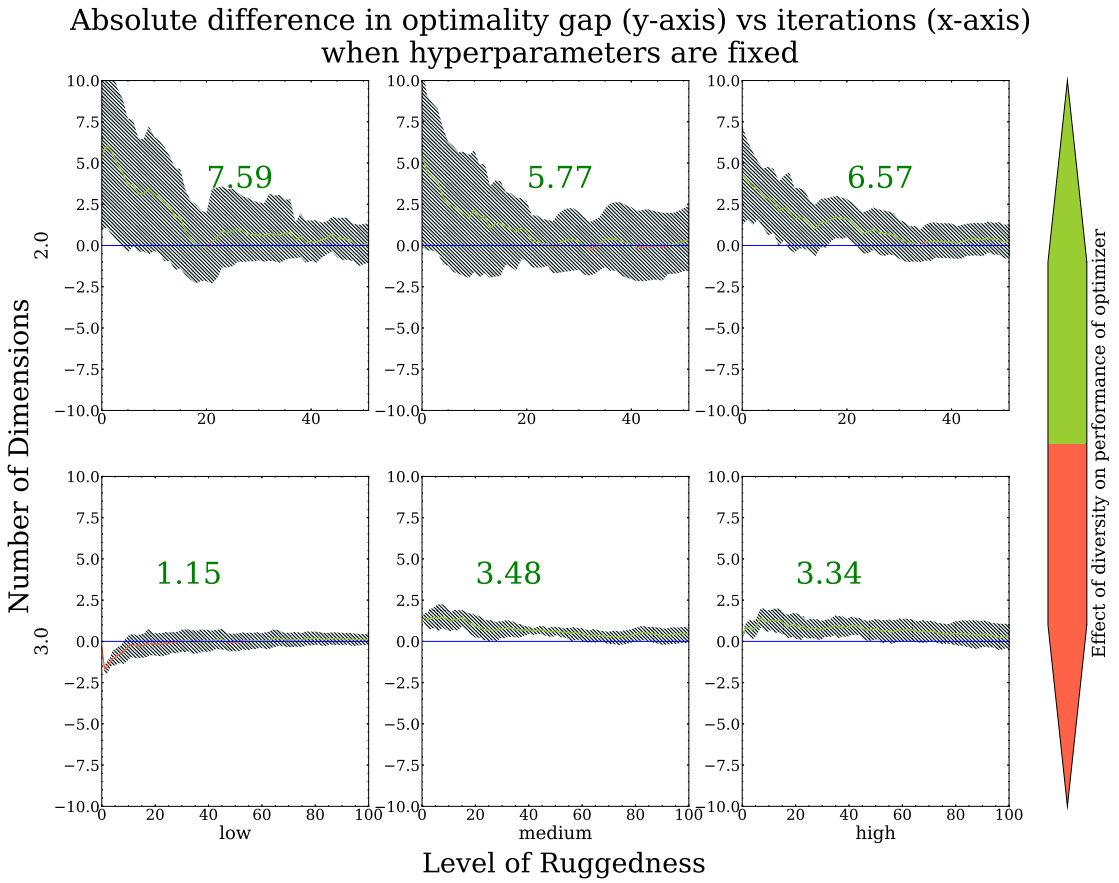


Figure 4.6: Optimality gap plot showing effects of diversity when the optimizer is not allowed to fit the hyper-parameters for the Gaussian Process and the hyper-parameters are instead fixed to the values found in Experiment A2. The results from this plot show positive NCOG values for all families of wildcat wells function even as dimensions increase, showing that once the 'Model Building advantage' is taken away the diverse samples outperform non-diverse samples. Further discussion on this plot can be read in Sec. 4.3

4.4 How does Diversity affect Optimization Convergence if the hyper-parameters are fixed to the optimal values?

This experiment is identical to our earlier study described at the beginning of the chapter, with two key differences: (1) we now fix the kernel hyper-parameters to the 'optimal hyper-parameter' values we found in Experiment 2 for all the instances in each family of the wildcat wells function, (2) and we do not allow either BO model to further optimize the kernel hyper-parameters. This should remove the hypothesized Model Building advantage of non-diverse

samples without altering any other aspects of Experiment 1 and the results in Fig. 4.1.

Figures 4.7 and 4.6 show that once the kernel hyper-parameters are fixed—removing the Model Building advantage of non-diverse samples—diverse samples consistently and robustly outperform non-diverse initial samples. This holds for both the initial Optimality Gap at the beginning of the search as well as the Cumulative Optimality Gap and is not qualitatively affected by the function smoothness or roughness amplitude. Unlike in Experiment 1 where diversity could either help or hurt the optimizer, once we remove the Model Building advantage, diversity only helps. This illustrates the causal effect that non-diverse samples are providing that accelerates BO convergence: they hasten the convergence of BO toward more accurate hyper-parameter estimates, which in turn accelerate optimization.

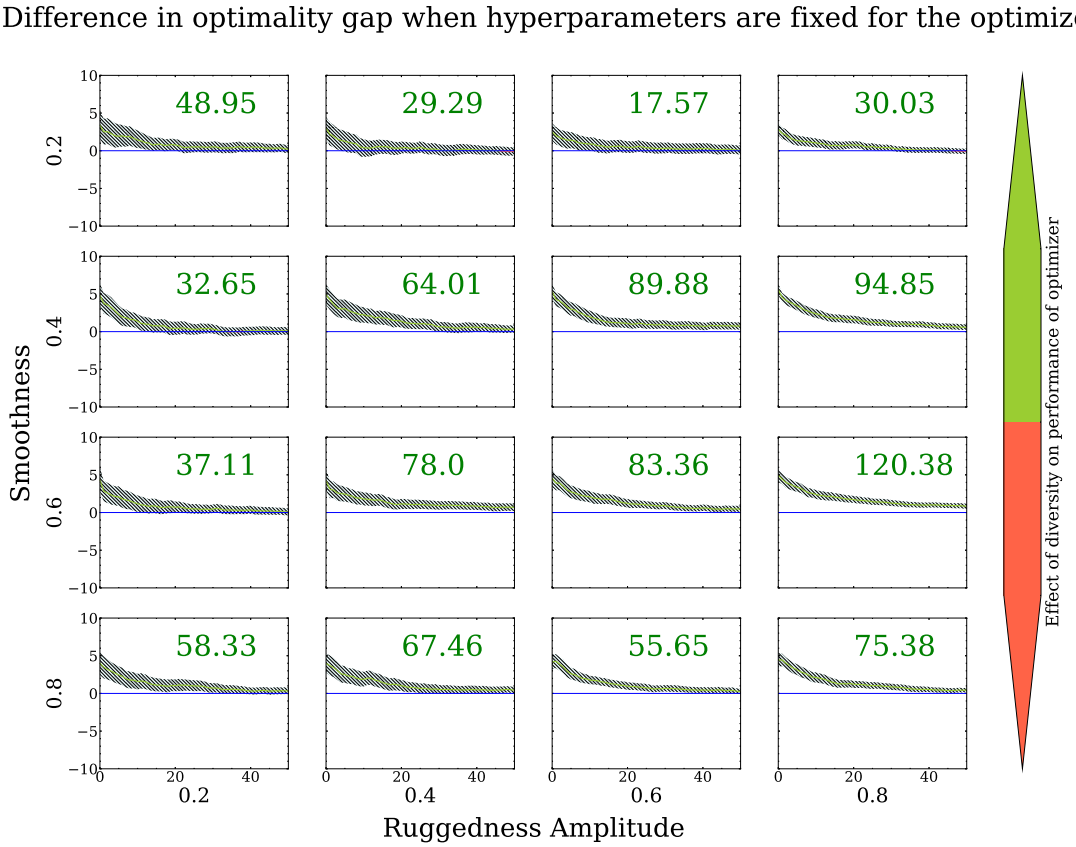


Figure 4.7: Experiment 3: Optimality gap plot showing effects of diversity when the optimizer is not allowed to fit the hyper-parameters for the Gaussian Process and the hyper-parameters are instead fixed to the values found in Experiment 2. The results from this plot show positive NCOG values for all families of wildcat wells function, showing that once the Model Building advantage' is taken away the diverse samples outperform non-diverse samples. Further discussion on this plot can be read in §4.4

Chapter 5: Extending Our Results to Other ND Functions

A natural question is whether our results are limited to just our choice of the wildcat-wells class of function generators, or do they transfer across different functions? To test this, we repeated the experiments described in Section 4 for three different but commonly used N-Dimensional optimization test functions: the Sphere, Rosenbrock and Rastrigin functions as seeen in Eq. 5.1. The only major difference with the previous experiments is that the difference plot for Rosenbrock is really zoomed out due to large values that the y-values can take in the function as seen in Fig. 5.3.

$$
\begin{aligned}
\text{Sphere}(X) &= \sum_{i=1}^{\text{dims}} x_i^2 \\
\text{Rastrigin}(X) &= 10 \times \text{dims} + \sum_{i=1}^{\text{dims}} \left[x_i^2 - 10\cos(2\pi x_i) \right] \\
\text{Rosenbrock}(X) &= \sum_{i=1}^{\text{dims-1}} \left[100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2 \right]
\end{aligned}
\tag{5.1}
$$

As seen in Fig. 5.1, when the hyperparameters are allowed to be optimized, in general low-diversity samples led to faster convergence than high-diversity initial samples. This is not always that case, as the 4D and 5D Rastrigin functions cases shows—in such cases non-diverse samples have comparatively marginal improvement in the longer term. For reference, this plot is designed

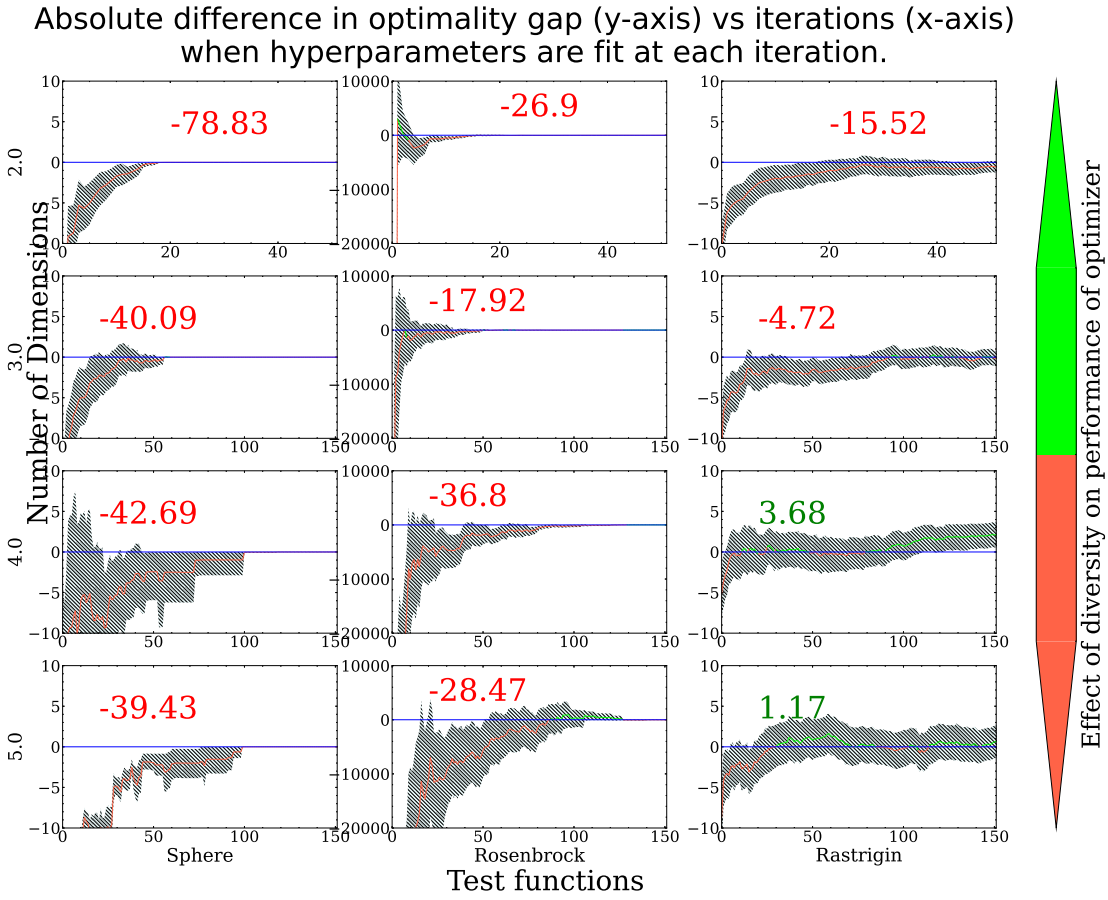


Figure 5.1: Optimality gap grid plot showing the absolute difference in current Optimality Gap between optimizers initialized with 5^{th} vs 95^{th} percentile diverse sample (y-axis) as a function of optimization iteration (x-axis). The different factors in the factor grid plot are the dimensions across the rows and the different test functions across the columns. Each plot also has text indicating the Percentage Cumulative Optimality Gap (PCOG), a positive value corresponds to a better performance by high diversity samples compared to the low diversity samples. The plot shows that BO benefits from diversity in some cases but not others. There are no obvious trends in how the PCOG values change in the grid. The results are further discussed in Sec. 5

to be a replication of study in Section 4, but just for different test functions.

Fig. 5.2 shows that 5^{th}-percentile diversity (low diversity) initial samples learns the kernel hyperparameter more accurately using fewer samples compared to 95^{th}-percentile diversity initial samples in two dimensions and that this holds true irrespective of the choice of test function. However, as the function dimension increases this effect diminishes since the number of initial samples needed to activate this "modeling advantage" regime increases (See earlier Fig. 4.4).

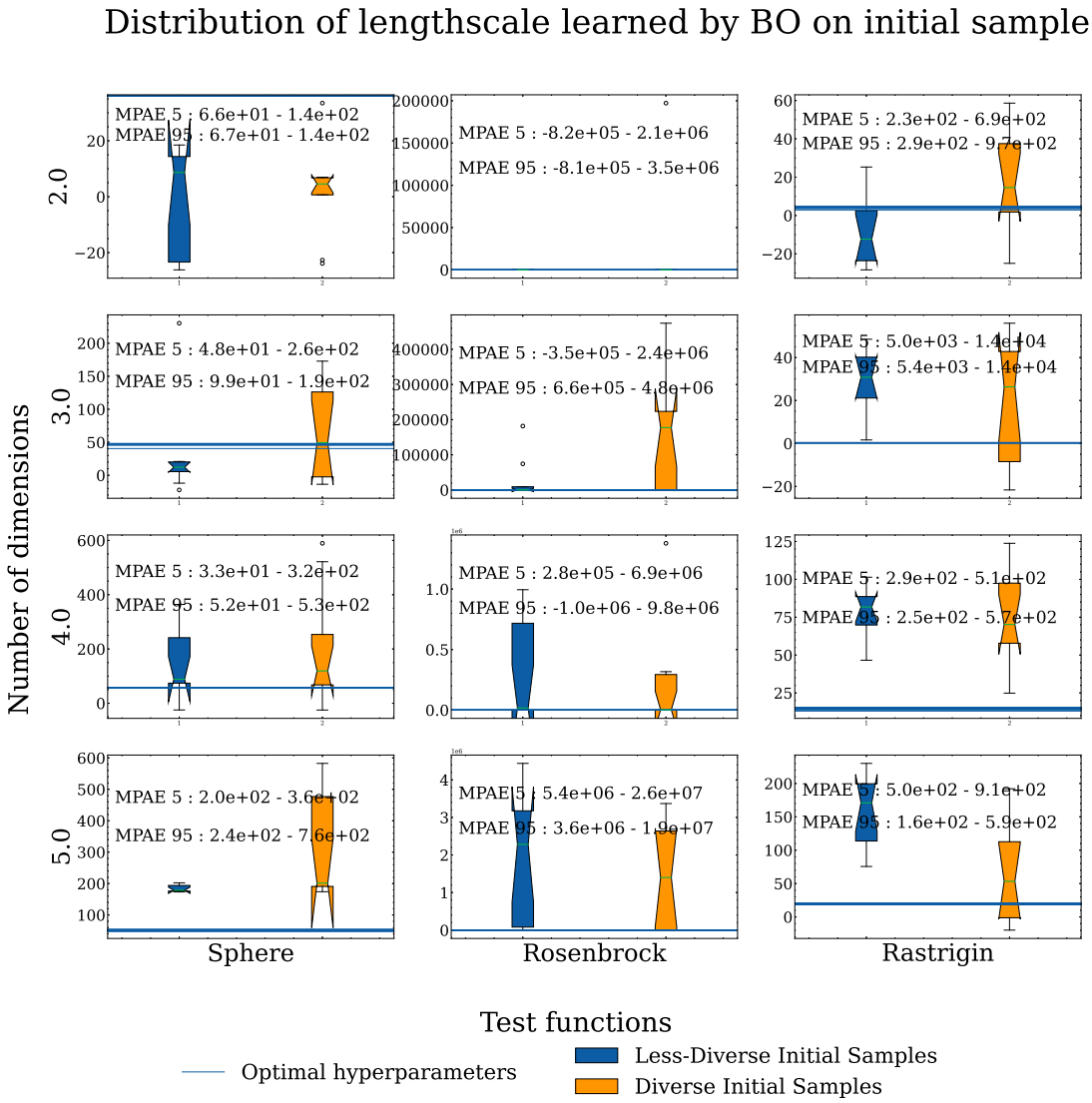


Figure 5.2: Box plot showing distribution of 'Lengthscale' hyper-parameter learned by BO when initiated with diverse (orange) and less-diverse samples (blue) for Sphere, Rosenbrock and Rastrigin test functions over 200 different seeds in each dimension. For refernce to how many training samples were used please check Table. 5.1. The optimal hyper-parameter for each test function over 10 different runs is also plotted as horizontal (blue) lines—in many but not all cases these overlap. Each cell in the plot also has the 95^{th} percentile confidence bound on Mean Absolute Error (MAE) for both diverse and non-diverse samples. The results show that MAE confidence bounds for non-diverse samples are smaller compared to diverse samples for most test functions but at least does as well as the 95^{th}. Thus, indicating a presence of Model Building advantage for non-diverse initial samples. The results of this figure are further discussed in Sec. 5

With this additional set of data, samples from from the 95^{th}-percentile of diversity learn the hyperparameters as well as 5^{th}-percentile samples. For reference, like with Fig. 4.5 above, this plot was designed to be a replication of our study in Sec. 4.2, but just for different test functions and

across increased dimensions. Unlike in Fig. 4.5 here we see that our proposed causal explanation for the "modeling advantage" is less clear since for certain functions the high-diversity samples have better posterior convergence than the 5^{th}-percentile samples, and vice versa depending on the specific function and dimension.

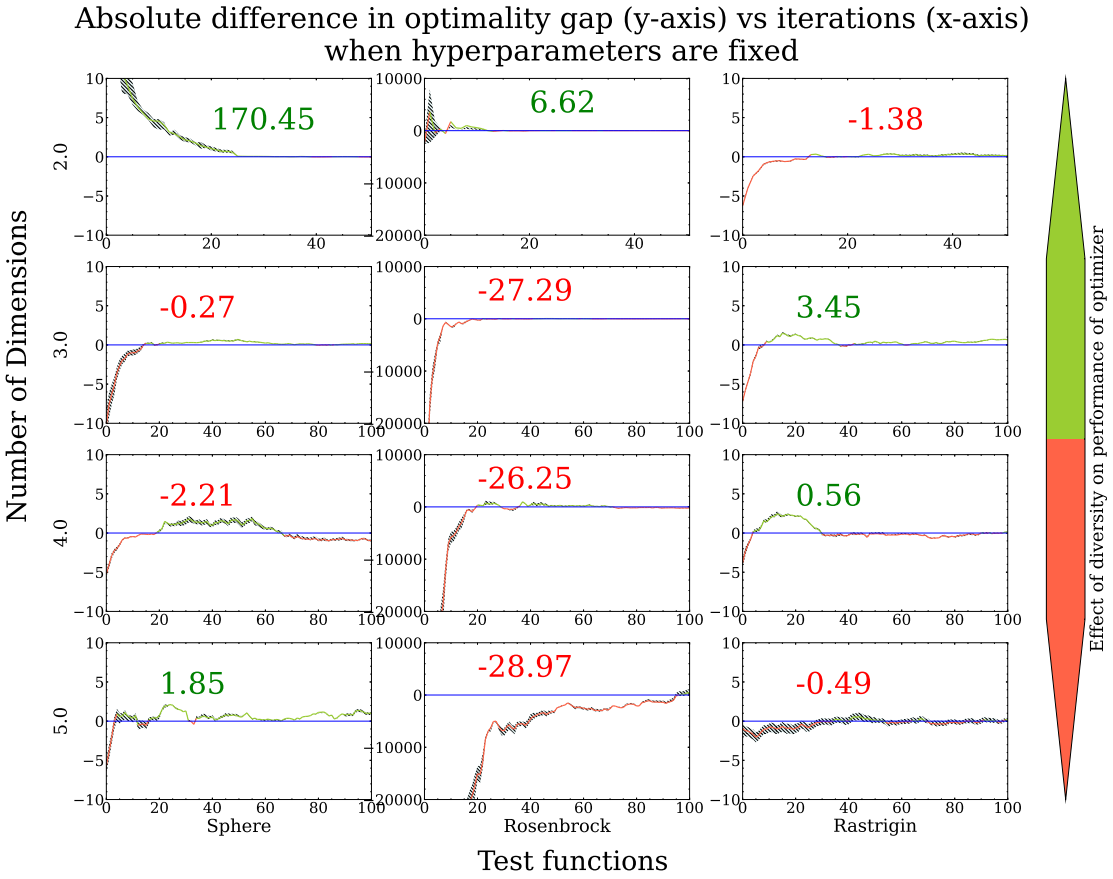


Figure 5.3: Optimality gap plot showing effects of diversity when the optimizer is not allowed to fit the hyper-parameters for the Gaussian Process and the hyper-parameters are instead fixed to the values found in Experiment A2. The results from this plot show signficantly improved PCOG values compared to Fig. 5.1. 'Rosenbrock' is the only test function that does not benefit from the diverse samples, its performance remains the same as it was when hyperparameters were optimized, Further discussion on this plot can be read in Sec. 5

In Fig. 5.3 where the kernel hyper-parameters are fixed to what should be optimal values, (compared to Fig. 5.1 where the kernel hyper-parameters are learned) we can see several effects. First, we see that the low diversity initial samples had, on average, better initial starting points on these test functions as seen by the PCOG values on the x-axis at "0". This could largely be

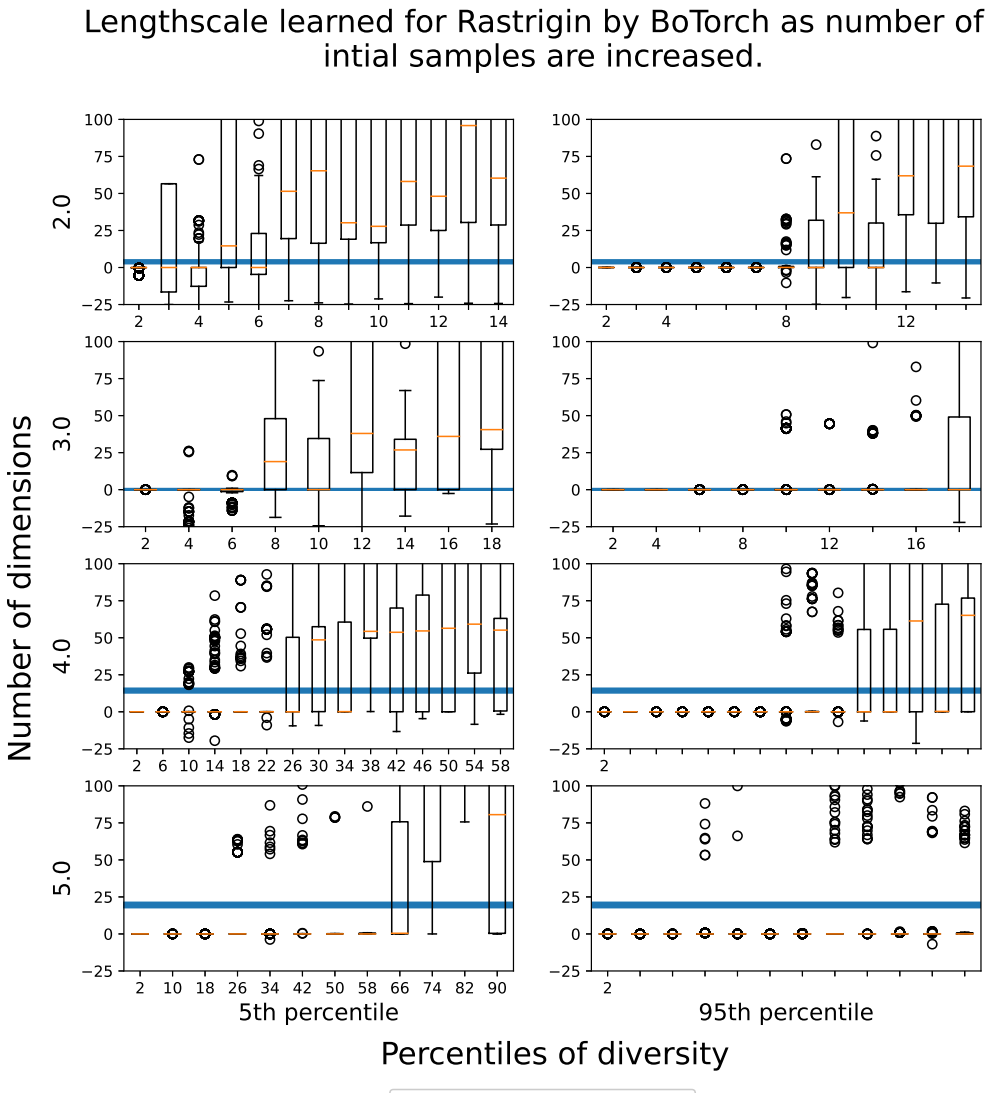


Figure 5.4: Box plot showing the lengthscale parameter as learned by Rastrigin test function in 2D and 3D as the training samples are increased. The plot also confirms the existence of a 'modeling advantage' for training samples of a particular size. The results are further discussed in Sec. 5

luck or a peculiarity with the three test functions, since common optimization test functions often have their optimal points toward the center of the domain, which non-diverse starting points are likely to sample with higher frequency compared to diverse starting points. (Note in our wildcat wells function this was not the case and the optimal point was likely to occur at any point in the domain depending on the seed of the random function generator.) Second, we see compared to Fig. 5.1 that high diversity initial samples appear to be able to benefit from the

'Space Exploration' advantage we hypothesized in Sec. 4 and do catch-up almost instantaneously compared to the lower-diversity samples. For reference, this plot is designed to be a replication of our study in Sec. 4.4, but just for different test functions. We still see a similar effect, in the sense that fixing the BO hyper-parameters aids the diverse initial sample condition, on average, which mirrors qualitatively the phenomenon we observed on the wildcat wells function (compare this supplemental material document's Fig. 5.1 with Fig. 5.3).

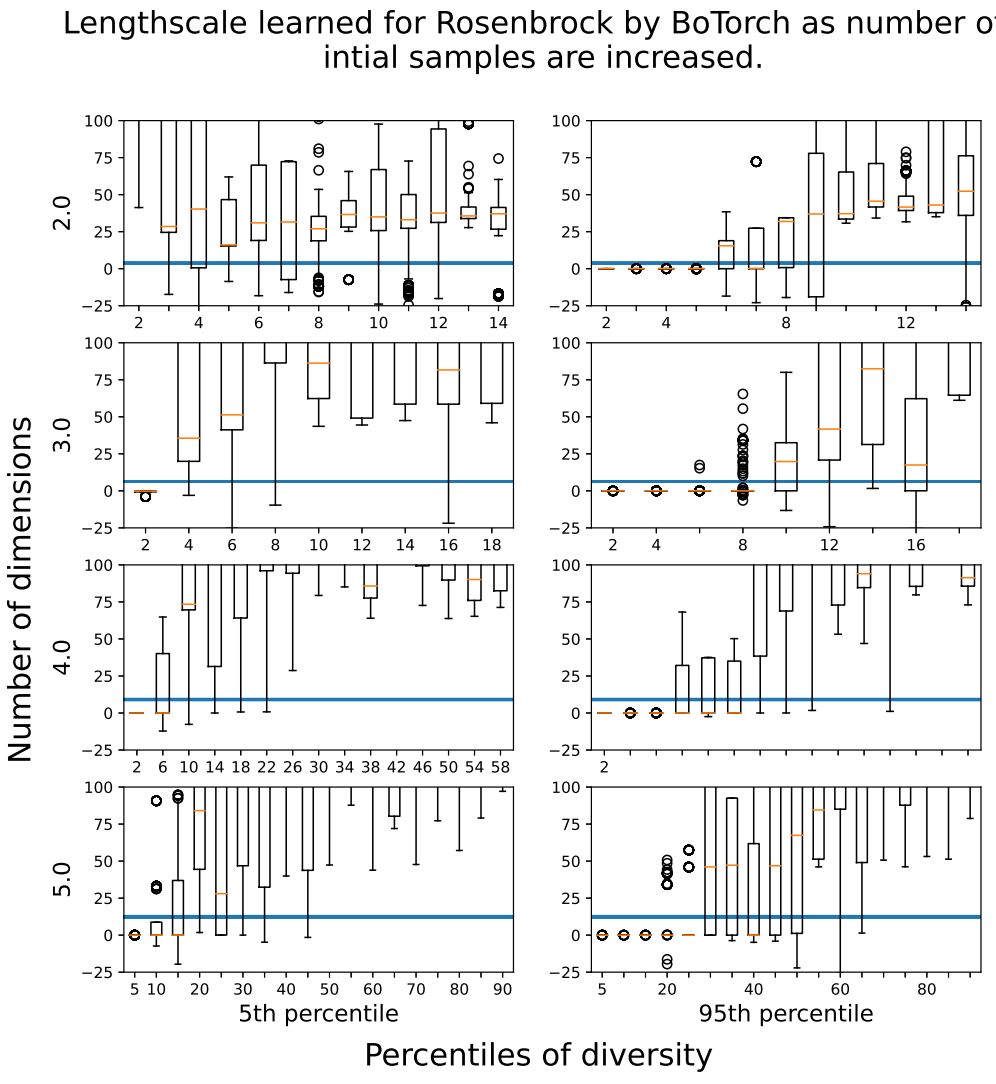


Figure 5.5: Box plot showing the lengthscale parameter as learned by Rosenbrock test function in 2D and 3D as the training samples are increased. The plot also confirms the existence of a 'modeling advantage' for training samples of a particular size. The results are further discussed in Sec. 5

In Figs. 5.4, 5.5, and 5.6 we can see how increasing the number of initial training samples induces convergence on the learned kernel hyper-parameters for the Rastrigin, Rosenbrock, and Sphere functions, respectively. We used these plots to choose the number of training samples to be used in Figs. 5.1, 5.2, and 5.3 by selecting the number of samples within the "model building advantage" regime (as opposed to the sample deficient or sample saturated regime). The specific number of training samples used for each function at each dimension can be seen in Table 5.1. We can see that the performance of high diversity samples is significantly better when compared to the performance in Fig. 5.1. The high diversity samples still struggle to improve performance for 'Rosenbrock' function, our hypothesis is that because the number of samples needed to learn the hyperparameters is exceedingly large for the Rosenbrock function (see Fig. 5.5) our proposed "modeling-advantage" is not that helpful to the optimizer, since it has already found a reasonable optimum by the time it has collected sufficient samples to converge to reasonable kernel estimates.

Dimension	Sphere	Rosenbrock	Rastrigin
2	8	4	5
3	12	5	7
4	38	8	30
5	75	20	60

Table 5.1: Table showing the different training size/number of examples used to initialize BO for different test functions in Figs. 5.1, 5.2, 5.3.

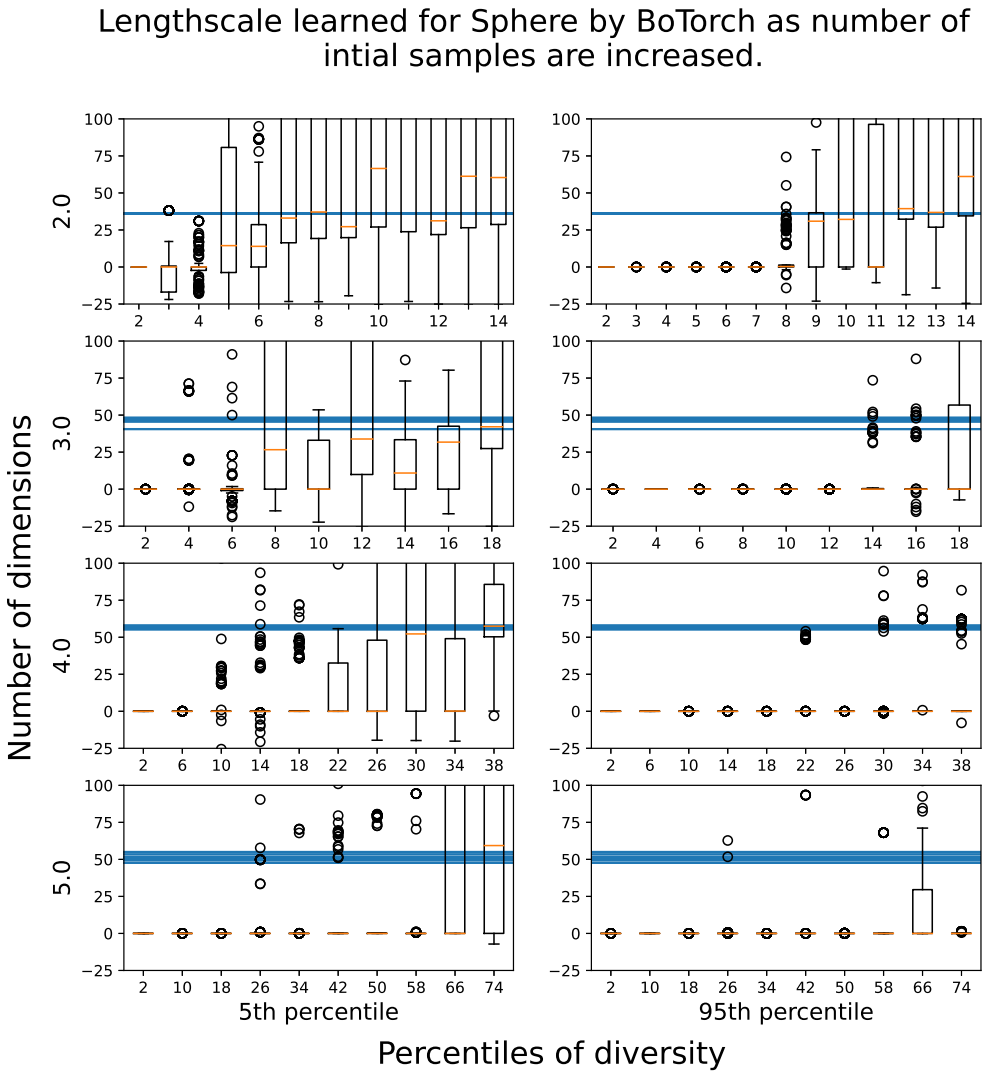


Figure 5.6: Box plot showing the lengthscale parameter as learned by Sphere test function in 2D and 3D as the training samples are increased. The plot also confirms the existence of a 'modeling advantage' for training samples of a particular size. The results are further discussed in Sec. 5

Chapter 6: Discussion and Conclusion

This thesis's original goal was to investigate how and when diverse initial samples help or hurt Bayesian Optimizers. Overall, we found that the initial diversity of the provided samples created two competing effects. First, Experiment 2 showed that non-diverse samples improved BO's abilities to quickly converge to optimal hyper-parameters—we called this a *Model Building* advantage. Second, Experiment 3 showed that conditioned on the same fixed hyper-parameters diverse samples improved BO's convergence to the optima through faster exploration of the space—we called this a *Space Exploration* advantage. In Experiment 1, diversity had mixed-to-negligible effects since both of these advantages were in play and competed with one another. This interaction provides insight for academic or industrial BO users since common practice recommends initializing BO with space-filling samples (to take advantage of the Space Exploration advantage), and ignores the Model Building advantage of non-diverse samples.

Beyond our main empirical result, our improvements to existing diverse sampling approaches (Sec. 2) provide new methods for studying how different percentile diversity sets affect phenomena. Researchers may find this contribution of separate technical and scientific interest for related studies that investigate the impact of diversity.

Beyond the individual results we observed and summarized in each experiment, there are some overall implications and limitations that may guide future work or interpretation of our

results more broadly, which we address below.

6.1 Where does this Model Building advantage induced by non-diverse samples come from?

As we conjectured in Experiment 2 (§4.2), and confirmed in Experiment 3 (§4.4), the key advantage of using non-diverse initial samples lies in their ability to induce faster and more accurate posterior convergence when inferring the optimal kernel hyper-parameters, such as length scale and others. This allowed the BO to make more judicious and aggressive choices about what points to sample next, so while the diversely initialized models might get a head start on exploring the space, non-diversely initialized models needed to explore less of the space overall, owing to tighter posteriors of possible functions under the Gaussian Process.

What this behavior implies more broadly is that non-diverse samples, whether given to an algorithm or a person, have a unique and perhaps underrated value in cases where we have high entropy priors over the Gaussian Process hyper-parameters or kernel. In such cases, sacrificing a few initial non-diverse points to better infer key length scales in the GP model may well be a worthwhile trade.

We also saw that in cases where the BO hyper-parameters were not further optimized (as in Experiment 3 where hyper-parameters were fixed to optimal values), using diverse points only helped BO. Researchers or practitioners using BO would benefit from carefully reviewing what kernel optimization strategy their library or implementation of choice actually does since that will affect whether or not the Model Building advantage of non-diverse samples is actually in play.

6.2 What if Hyper-parameters are fixed to non-optimal values?

We showed in Experiment 3 that fixing BO hyper-parameters to their optimal values ahead of time using an oracle allowed diverse initial samples to unilaterally outperform non-diverse samples. An interesting avenue of future work that we did not explore here for scope reasons would be to see if this holds when hyper-parameters are fixed to non-optimal values. In practical problems, we will not often know the optimal hyper-parameters ahead of time as we did in Experiment 3 which caused diversity's unilateral advantage, so we do not have evidence to generalize beyond this. However, our explanation of the Model Building advantage would predict that, so long as the hyper-parameters remain fixed (to any value), BO would not have a practical mechanism to benefit much from non-diverse samples, on average.

6.3 What are the implications for how we currently initialize BO?

One of our result's most striking implications is how it might influence BO initialization procedures that are often considered standard practice. For example, it is common to initialize a BO procedure with a small number of initial space-filling designs, using techniques like Latin Hypercube Sampling (LHS) before allowing BO to optimize its acquisition function for future samples. In cases where the BO hyper-parameters will remain fixed, Experiment 3 implies that this standard practice is excellent advice and far better than non-diverse samples. However, in cases where you plan to optimize the BO kernel during search, using something like LHS becomes more suspect.

In principle, from Experiment 1 we see that diverse samples may help or hurt BO, depend-

ing on how much leverage the Model Building advantage of the non-diverse samples can provide. For example, in the upper right of Fig. 4.1 the function is effectively random noise, and so there is not a strong Model Building advantage to be gained. In contrast, in the lower left, the smooth and well-behaved functions allowed non-diverse initialization to gain an upper hand.

Our results propose a perhaps now obvious initialization strategy: if you plan on optimizing the BO hyper-parameters, use some non-diverse samples to strategically provide an early Model Building advantage, while leveraging the rest of the samples to diversely cover the space.

6.4 How might other acquisition functions modulate diversity's effect?

While we have been referring to BO as though it is a single method throughout this thesis, individual BO implementations can vary, both in terms of their kernel structure and their choice of acquisition function—or how BO uses information about the underlying fitted Gaussian Process to select subsequent points. In the studies in Chapter 4 and 5, we used Expected Improvement (EI) since it is one of the most widespread choices, and behaves qualitatively like other common improvement-based measures like Probability of Improvement, Posterior Mean, and Upper Confidence Bound functions. Indeed, we hypothesize that part of the reason non-diverse initial samples are able to gain a Model Building advantage over diverse samples is due to a faster collapse in the posterior distribution of possible GP functions which serves as strong input to EI methods and related variants.

Yet EI and its cousins are only one class of acquisition function; would our results hold if we were to pick an acquisition function that directly attacked the GP's posterior variance? For example, either Entropy-based or Active Learning based acquisition functions? This thesis

did not test this and it would be a logical and valuable future study. Our experimental results and proposed explanation would predict the following: the Model Building advantage seen by non-diverse samples should reduce or disappear in cases where the acquisition function explicitly samples new points to minimize the posterior over GP function classes since in such cases BO itself would try to select samples that reduced overall GP variance, reducing its dependence on what the initial samples provide.

6.5 To what extent should we expect these results to generalize to other types of problems?

We selected a simple 2D function with controllable complexity in this thesis to aid in experimental simplicity, speed, replicability, and ease of visualization; however, this does raise the question of whether or not these results would truly transfer to more complex problems of engineering interest. Chapter 5 addressed additional common optimization test functions with different properties, though it is impossible to claim in general that the phenomena studied by this thesis would extend to *every* design problem. While future work would have to address a larger class of more complex problems, we can look at a few critical problem-specific factors and ask what our proposed explanatory model would predict.

For higher dimensional problems, standard GP kernel choices like RBF or Matérn begin to face exponential cost increases due to how hyper-volumes expand. In such cases, having strong constraints (via hyper-parameter priors or posteriors) over possible GP functions becomes increasingly important for fast BO convergence. Our results would posit that any Model Building advantages from non-diverse sampling would become increasingly important or impactful in

cases where it helped BO rapidly collapse the hyper-parameter posteriors.

For discontinuous functions (or GP kernels that assumed as much), the Model Building advantage of non-diverse samples would decrease since large sudden jumps in the GP posterior mean and variance would make it harder for BO to exploit a Model Building advantage. However, in discontinuous cases where there were still common global smoothness parameters that governed the continuous portions the Model Building advantage would still accelerate advantages for BO convergence.

6.6 How might the results guide human subject experiments or understanding of human designers?

One possible implication of our results for human designers is that the effects of example diversity on design outcomes may vary as a function of designer's prior knowledge of the design problem. More specifically, the Model Building advantage observed in Experiment 2 (and subsequent removal in Experiment 3) suggests that when designers have prior knowledge of how quickly the function changes in a local area of the design space, they can more reliably benefit from the Space Exploration advantage of diverse examples. This leads to a potentially counter-intuitive prediction that domain experts may benefit more from diverse examples compared to domain novices since domain experts would tend to have prior knowledge of the nature of the design problem (a Model Building advantage). Additionally, perhaps under conditions of uncertainty about the nature of the design problem, it would be useful to combine the strengths of diverse and non-diverse examples; this could be accomplished with a cluster-sampling approach, where we sample diverse points of the design space, but include local non-diverse clusters of

examples that are nearby, to facilitate learning of the shape of the design function.

Bibliography

[1] Katherine Fu, Joel Chan, Jonathan Cagan, Kenneth Kotovsky, Christian Schunn, and Kristin Wood. The Meaning of "Near" and "Far": The Impact of Structuring Design Databases and the Effect of Distance of Analogy on Design Output. *Journal of Mechanical Design*, 135 (2), January 2013. ISSN 1050-0472. doi: 10.1115/1.4023158. URL https://doi.org/10.1115/1.4023158.

[2] Joel Chan, Katherine Fu, Christian Schunn, Jonathan Cagan, Kristin Wood, and Kenneth Kotovsky. On the Benefits and Pitfalls of Analogies for Innovative Design: Ideation Performance Based on Analogical Distance, Commonness, and Modality of Examples. *Journal of Mechanical Design*, 133(8), August 2011. ISSN 1050-0472. doi: 10.1115/1.4004396. URL https://doi.org/10.1115/1.4004396.

[3] Suzana Linic, Vojkan Lucanin, Srdjan Zivkovic, Marko Rakovic, and Mirjana Puharic. Experimental and Numerical Methods for Concept Design and Flow Transition Prediction on the Example of the Bionic High-Speed Train. In Nenad Mitrovic, Goran Mladenovic, and Aleksandra Mitrovic, editors, *Experimental and Computational Investigations in Engineering*, Lecture Notes in Networks and Systems, pages 65–82, Cham, 2021. Springer International Publishing. ISBN 978-3-030-58362-0. doi: 10.1007/978-3-030-58362-0_5.

[4] Diana P. Moreno, Luciënne T. Blessing, Maria C. Yang, Alberto A. Hernández, and Kristin L. Wood. Overcoming design fixation: Design by analogy studies and nonintuitive findings. *AI EDAM*, 30(2):185–199, May 2016. ISSN 0890-0604, 1469-1760. doi: 10.1017/S0890060416000068. URL https://www.cambridge.org/core/journals/ai-edam/article/overcoming-design-fixation-design-by-analogy-studies-and-nonintuitive-038760839BBBEFC08F146457A77BBE51. Publisher: Cambridge University Press.

[5] Jonali Baruah and Paul B. Paulus. Category assignment and relatedness in the group ideation process. *Journal of Experimental Social Psychology*, 47(6):1070–1077, November 2011. ISSN 0022-1031. doi: 10.1016/j.jesp.2011.04.007. URL https://www.sciencedirect.com/science/article/pii/S0022103111001156.

[6] Pao Siangliulue, Kenneth C. Arnold, Krzysztof Z. Gajos, and Steven P. Dow. Toward Collaborative Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 937–945, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-2922-4. doi: 10.1145/2675133.2675239. URL http://doi.acm.org/10.1145/2675133.2675239.

[7] Bernard A Nijstad, Wolfgang Stroebe, and Hein F. M Lodewijkx. Cognitive stimulation and interference in groups: Exposure effects in an idea generation task. *Journal of Experimental Social Psychology*, 38(6):535–544, November 2002. ISSN 0022-1031. doi: 10.1016/S0022-1031(02)00500-0. URL https://www.sciencedirect.com/science/article/pii/S0022103102005000.

[8] A. Taylor and H. R. Greve. Superman or the Fantastic Four? Knowledge Combination and Experience in Innovative Teams. *Academy of Management Journal*, 49(4):723–740, 2006. ISSN 0001-4273.

[9] Liang Zeng, Robert W. Proctor, and Gavriel Salvendy. Fostering creativity in product and service development: validation in the domain of information technology. *Human Factors*, 53(3):245–270, June 2011. ISSN 0018-7208. doi: 10.1177/0018720811409219.

[10] Paul A. Howard-Jones, Sarah-Jayne . J. Blakemore, Elspeth A. Samuel, Ian R. Summers, and Guy Claxton. Semantic divergence and creative story generation: An fMRI investigation. *Cognitive Brain Research*, 25(1):240–250, 2005.

[11] Joel Chan and Christian D. Schunn. The importance of iteration in creative conceptual combination. *Cognition*, 145:104–115, December 2015. ISSN 0010-0277. doi: 10.1016/j.cognition.2015.08.008. URL http://www.sciencedirect.com/science/article/pii/S0010027715300524.

[12] M. M. Gielnik, M. Frese, J. M. Graf, and A. Kampschulte. Creativity in the opportunity identification process and the moderating effect of diversity of information. *Journal of Business Venturing*, 27(5):559–576, 2011. ISSN 08839026. doi: 10.1016/j.jbusvent.2011.10.003.

[13] Niek Althuizen and Berend Wierenga. Supporting Creative Problem Solving with a Case-Based Reasoning System. *Journal of Management Information Systems*, 31(1):309–340, 2014. ISSN 0742-1222. doi: 10.2753/MIS0742-1222310112.

[14] Huan Yuan, Kelong Lu, Mengsi Jing, Cuirong Yang, and Ning Hao. Examples in creative exhaustion: The role of example features and individual differences in creativity. *Personality and Individual Differences*, 189:111473, April 2022. ISSN 0191-8869. doi: 10.1016/j.paid.2021.111473. URL https://www.sciencedirect.com/science/article/pii/S0191886921008527.

[15] Alex Doboli, Anurag Umbarkar, Varun Subramanian, and Simona Doboli. Two experimental studies on creative concept combinations in modular design of electronic embedded systems. *Design Studies*, 35(1):80–109, 2014. ISSN 0142694X. doi: 10.1016/j.destud.2013.10.002.

[16] Sunhee Jang. The Effect of Image Stimulus on Conceptual Combination in the Design Idea Generation Process. *Archives of Design Research*, 112(4):59, 2014. ISSN 1226-8046. doi: 10.15187/adr.2014.11.112.4.59.

[17] Michele I. Mobley, Lesli M. Doares, and Michael D. Mumford. Process analytic models of creative capacities: Evidence for the combination and reorganization process. *Creativity Research Journal*, 5(2), 1992. ISSN 1532-6934. doi: 10.1080/10400419209534428. Place: US Publisher: Lawrence Erlbaum.

[18] Wayne A. Baughman and Michael D. Mumford. Process-Analytic Models of creative Capacities: Operations Influencing the Combination-and-Reorganization Process. *Creativity Research Journal*, 8(1):37–62, January 1995. ISSN 1040-0419. doi: 10.1207/s15326934crj0801_4. URL https://doi.org/10.1207/s15326934crj0801_4. Publisher: Routledge _eprint: https://doi.org/10.1207/s15326934crj0801_4.

[19] Chandrika Kamath. Intelligent Sampling for Surrogate Modeling, Hyperparameter Optimization, and Data Analysis. Technical Report LLNL-TR-829837, Lawrence Livermore National Lab. (LLNL), Livermore, CA (United States), December 2021. URL https://www.osti.gov/biblio/1836193.

[20] Zhongkun Ma and Guy A. E. Vandenbosch. Impact of Random Number Generators on the performance of particle swarm optimization in antenna design. In *2012 6th European Conference on Antennas and Propagation (EUCAP)*, pages 925–929, Prague, Czech Republic, March 2012. IEEE. doi: 10.1109/EuCAP.2012.6205998.

[21] Borhan Kazimipour, Xiaodong Li, and A. K. Qin. Effects of population initialization on differential evolution for large scale optimization. In *2014 IEEE Congress on Evolutionary Computation (CEC)*, pages 2404–2411, Beijing, China, July 2014. doi: 10.1109/CEC.2014.6900624. ISSN: 1941-0026.

[22] H. Maaranen, K. Miettinen, and M. M. Mäkelä. Quasi-random initial population for genetic algorithms. *Computers & Mathematics with Applications*, 47(12):1885–1895, June 2004. ISSN 0898-1221. doi: 10.1016/j.camwa.2003.07.011. URL https://www.sciencedirect.com/science/article/pii/S0898122104840240.

[23] Xin-She Yang. Swarm intelligence based algorithms: a critical analysis. *Evol. Intel.*, 7(1):17–28, April 2014. ISSN 1864-5917. doi: 10.1007/s12065-013-0102-2. URL https://doi.org/10.1007/s12065-013-0102-2.

[24] Qian Li, San-Yang Liu, and Xin-She Yang. Influence of initialization on the performance of metaheuristic optimizers. *Applied Soft Computing*, 91:106193, June 2020. ISSN 1568-4946. doi: 10.1016/j.asoc.2020.106193. URL https://www.sciencedirect.com/science/article/pii/S1568494620301332.

[25] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, April 2002. ISSN 1941-0026. doi: 10.1109/4235.996017. Conference Name: IEEE Transactions on Evolutionary Computation.

[26] Leshi Shu, Ping Jiang, Xinyu Shao, and Yan Wang. A New Multi-Objective Bayesian Optimization Formulation With the Acquisition Function for Convergence and Diversity.

Journal of Mechanical Design, 142(9), March 2020. ISSN 1050-0472. doi: 10.1115/1.4046508. URL https://doi.org/10.1115/1.4046508.

[27] Alex Kulesza and Ben Taskar. Determinantal Point Processes for Machine Learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, December 2012. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000044. URL https://www.nowpublishers.com/article/Details/MAL-044. Publisher: Now Publishers, Inc.

[28] Alex Kulesza and Ben Taskar. k-DPPs: fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 1193–1200, Bellevue, Washington, USA, June 2011. Omnipress. ISBN 978-1-4503-0619-5.

[29] Eesh Kamrah, Fatemeh Ghoreishi, Zijian Ding, Joel Chan, and Mark Fuge. How diverse initial samples help and hurt bayesian optimizers. *Journal of Mechanical Design*, pages 1–32, 2023. doi: https://doi.org/10.1115/1.4063006.

[30] T.W. Simpson, J.D. Poplinski, P. N. Koch, and J.K. Allen. Metamodels for Computer-based Engineering Design: Survey and recommendations. *Engineering with Computers*, 17(2):129–150, July 2001. ISSN 1435-5663. doi: 10.1007/PL00007198. URL https://doi.org/10.1007/PL00007198.

[31] Nestor V. Queipo, Raphael T. Haftka, Wei Shyy, Tushar Goel, Rajkumar Vaidyanathan, and P. Kevin Tucker. Surrogate-based analysis and optimization. *Progress in Aerospace Sciences*, 41(1):1–28, January 2005. ISSN 0376-0421. doi: 10.1016/j.paerosci.2005.02.001. URL https://www.sciencedirect.com/science/article/pii/S0376042105000102.

[32] R. Jin, W. Chen, and T.W. Simpson. Comparative studies of metamodelling techniques under multiple modelling criteria. *Structural and Multidisciplinary Optimization*, 23(1):1–13, December 2001. ISSN 1615-1488. doi: 10.1007/s00158-001-0160-4. URL https://doi.org/10.1007/s00158-001-0160-4.

[33] Thurston Sexton and Max Yi Ren. Learning an Optimization Algorithm Through Human Design Iterations. *Journal of Mechanical Design*, 139(10):10, October 2017. ISSN 1050-0472. doi: 10.1115/1.4037344. URL https://doi.org/10.1115/1.4037344.

[34] Sean Tauber, Daniel J. Navarro, Amy Perfors, and Mark Steyvers. Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review*, 124(4):410–441, July 2017. ISSN 1939-1471, 0033-295X. doi: 10.1037/rev0000052. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/rev0000052.

[35] Charles Kemp and Joshua B. Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, August 2008. doi: 10.1073/pnas.0802631105. URL https://www.pnas.org/doi/10.1073/pnas.0802631105. Publisher: Proceedings of the National Academy of Sciences.

[36] Hongjing Lu, Alan L. Yuille, Mimi Liljeholm, Patricia W. Cheng, and Keith J. Holyoak. Bayesian generic priors for causal learning. *Psychological Review*, 115(4):955–984, 2008. ISSN 1939-1471. doi: 10.1037/a0013256.

[37] Hongjing Lu, Dawn Chen, and Keith J. Holyoak. Bayesian analogy with relational transformations. *Psychological Review*, 119(3):617–648, July 2012. ISSN 1939-1471. doi: 10.1037/a0028719.

[38] Paul Saves, Nathalie Bartoli, Youssef Diouane, Thierry Lefebvre, Joseph Morlier, Christophe David, Eric Nguyen Van, and Sébastien Defoort. Bayesian optimization for mixed variables using an adaptive dimension reduction process: applications to aircraft design. In *AIAA SCITECH 2022 Forum*, AIAA SciTech Forum. American Institute of Aeronautics and Astronautics, December 2021. doi: 10.2514/6.2022-0082. URL https://arc.aiaa.org/doi/10.2514/6.2022-0082.

[39] Chunfeng Cui, Tao Ouyang, Chao Tang, Chaoyu He, Jin Li, Chunxiao Zhang, and Jianxin Zhong. Bayesian optimization-based design of defect gamma-graphyne nanoribbons with high thermoelectric conversion efficiency. *Carbon*, 176:52–60, May 2021. ISSN 0008-6223. doi: 10.1016/j.carbon.2021.01.126. URL https://www.sciencedirect.com/science/article/pii/S0008622321001469.

[40] Michael Hellwig and Hans-Georg Beyer. Benchmarking evolutionary algorithms for single objective real-valued constrained optimization – A critical review. *Swarm and Evolutionary Computation*, 44:927–944, February 2019. ISSN 2210-6502. doi: 10.1016/j.swevo.2018.10.002. URL https://www.sciencedirect.com/science/article/pii/S2210650218305406.

[41] Danial Yazdani, Mohammad Nabi Omidvar, Ran Cheng, Jürgen Branke, Trung Thanh Nguyen, and Xin Yao. Benchmarking Continuous Dynamic Optimization: Survey and Generalized Test Suite. *IEEE Transactions on Cybernetics*, 52(5):3380–3393, May 2022. ISSN 2168-2275. doi: 10.1109/TCYB.2020.3011828. Conference Name: IEEE Transactions on Cybernetics.

[42] Ali Ahrari, Saber Elsayed, Ruhul Sarker, Daryl Essam, and Carlos A. Coello Coello. A Novel Parametric benchmark generator for dynamic multimodal optimization. *Swarm and Evolutionary Computation*, 65:100924, August 2021. ISSN 2210-6502. doi: 10.1016/j.swevo.2021.100924. URL https://www.sciencedirect.com/science/article/pii/S2210650221000857.

[43] Mark Fuge, Josh Stroud, and Alice Agogino. Automatically Inferring Metrics for Design Creativity. In *IDETC-CIE2013*, volume Volume 5: 25th International Conference on Design Theory and Methodology; ASME 2013 Power Transmission and Gearing Conference, Portland, Oregon, USA, August 2013. American Society of Mechanical Engineers (ASME). doi: 10.1115/DETC2013-12620. URL https://doi.org/10.1115/DETC2013-12620. V005T06A010.

[44] Faez Ahmed, Sharath Kumar Ramachandran, Mark Fuge, Sam Hunter, and Scarlett Miller. Design Variety Measurement Using Sharma–Mittal Entropy. *Journal of Mechanical Design*, 143(6):14, June 2021. ISSN 1050-0472. doi: 10.1115/1.4048743. URL https://doi.org/10.1115/1.4048743. 061702.

[45] Scarlett R. Miller, Samuel T. Hunter, Elizabeth Starkey, Sharath Ramachandran, Faez Ahmed, and Mark Fuge. How Should We Measure Creativity in Engineering Design? A Comparison Between Social Science and Engineering Approaches. *Journal of Mechanical Design*, 143(3), March 2021. ISSN 1050-0472. doi: 10.1115/1.4049061. URL https://doi.org/10.1115/1.4049061.

[46] Faez Ahmed, Sharath Kumar Ramachandran, Mark Fuge, Samuel Hunter, and Scarlett Miller. Interpreting Idea Maps: Pairwise Comparisons Reveal What Makes Ideas Novel. *Journal of Mechanical Design*, 141(2):13, February 2019. ISSN 1050-0472. doi: 10.1115/1.4041856. URL https://doi.org/10.1115/1.4041856.

[47] Faez Ahmed and Mark Fuge. Ranking Ideas for Diversity and Quality. *Journal of Mechanical Design*, 140(1):11, January 2018. ISSN 1050-0472. doi: 10.1115/1.4038070. URL https://doi.org/10.1115/1.4038070.

[48] Faez Ahmed. *Diversity and Novelty: Measurement, Learning and Optimization*. PhD Thesis, University of Maryland, College Park MD, 2019. URL http://hdl.handle.net/1903/25383.

[49] Chao Li, Xiaogeng Chu, Yingwu Chen, and Lining Xing. A knowledge-based initialization technique of genetic algorithm for the travelling salesman problem. In *2015 11th International Conference on Natural Computation (ICNC)*, pages 188–193, August 2015. doi: 10.1109/ICNC.2015.7377988. ISSN: 2157-9563.

[50] Na Dong, Chun-Ho Wu, Wai-Hung Ip, Zeng-Qiang Chen, Ching-Yuen Chan, and Kai-Leung Yung. An opposition-based chaotic GA/PSO hybrid algorithm and its application in circle detection. *Computers & Mathematics with Applications*, 64(6):1886–1902, September 2012. ISSN 0898-1221. doi: 10.1016/j.camwa.2012.03.040. URL https://www.sciencedirect.com/science/article/pii/S0898122112002453.

[51] Hadi Eskandar, Ali Sadollah, Ardeshir Bahreininejad, and Mohd Hamdi. Water cycle algorithm – A novel metaheuristic optimization method for solving constrained engineering optimization problems. *Computers & Structures*, 110-111:151–166, November 2012. ISSN 0045-7949. doi: 10.1016/j.compstruc.2012.07.010. URL https://www.sciencedirect.com/science/article/pii/S0045794912001770.

[52] Dmytro Mishkin and Jiri Matas. All you need is a good init, February 2016. URL http://arxiv.org/abs/1511.06422. arXiv:1511.06422 [cs].

[53] Weiwei Yuan, Yongkoo Han, Donghai Guan, Sungyoung Lee, and Young-Koo Lee. Initial training data selection for active learning. In *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication*, ICUIMC '11, pages 1–7, New York, NY, USA, February 2011. Association for Computing Machinery. ISBN

978-1-4503-0571-6. doi: 10.1145/1968613.1968619. URL https://doi.org/10.1145/1968613.1968619.

[54] Burr Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer International Publishing, Cham, 2012. ISBN 978-3-031-00432-2 978-3-031-01560-1. doi: 10.1007/978-3-031-01560-1. URL https://link.springer.com/10.1007/978-3-031-01560-1.

[55] Jinsung Yoon, Sercan Arik, and Tomas Pfister. Data Valuation using Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10842–10851. PMLR, November 2020. URL https://proceedings.mlr.press/v119/yoon20a.html.

[56] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is All You Need: Learning Skills without a Reward Function. In *ICLR 2019*, New Orleans, Louisiana, United States, December 2018. URL https://openreview.net/forum?id=SJx63jRqFm.

[57] W. Mason and D. J. Watts. Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, 109(3):764–769, January 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1110069108. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1110069108.

[58] Maximilian Balandat, Brian Karrer, Daniel Jiang, Samuel Daulton, Ben Letham, Andrew G Wilson, and Eytan Bakshy. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems*, volume 33, pages 21524–21538. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f5b1b89d98b7286673128a5fb112cb9a-Abstract.html.

[59] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, SciPy 1.0 Contributors, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith,

Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17 (3):261–272, March 2020. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0686-2. URL http://www.nature.com/articles/s41592-019-0686-2.

[60] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. June 2018. doi: 10.7551/mitpress/4175.001.0001. URL https://direct.mit.edu/books/book/1821/Learning-with-KernelsSupport-Vector-Machines.

[61] Daniele Calandriello, Michal Derezinski, and Michal Valko. Sampling from a k-DPP without looking at all items. In *Advances in Neural Information Processing Systems*, volume 33, pages 6889–6899. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4d410063822cd9be28f86701c0bc3a31-Abstract.html.

[62] Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Efficient Sampling for k-Determinantal Point Processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 1328–1337. PMLR, May 2016. URL https://proceedings.mlr.press/v51/li16f.html. ISSN: 1938-7228.

[63] Zelda Elaine Mariet. Learning and enforcing diversity with Determinantal Point Processes. Master's thesis, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 2016. URL https://dspace.mit.edu/bitstream/handle/1721.1/103671/953457802-MIT.pdf.

[64] R. J. Serfling. Probability Inequalities for the Sum in Sampling without Replacement. *The Annals of Statistics*, 2(1):39–48, January 1974. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176342611. URL https://projecteuclid.org/journals/annals-of-statistics/volume-2/issue-1/Probability-Inequalities-for-the-Sum-in-Sampling-without-Replacement/10.1214/aos/1176342611.full.