Dimensionality Reduction Techniques for Global Optimization

Adilet Otemissov
Exeter College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy in Mathematics

Michaelmas 2020

I seek refuge in Allah (God) from Satan the accused.
In the name of Allah, the Most Gracious, the Most Merciful.

I begin by praising Allah (God), my Master and my Creator, who guided me to the right path when I was astray and who granted me beneficial knowledge that many do not possess and who provides me with such sustenance that many are deprived of. I thank him for this life, for the guidance, for the knowledge, for the wealth and for many other blessings that I am aware of and that I am not aware of. All Praise and Thanks belong to Allah the Most Gracious, the Most Merciful, the Creator of this universe and everything in it.

Then, I send my salutations upon prophet Muhammad — my teacher, through whom Holy Quran was revealed and illuminated with its light the Earth when it was darkest, and watered lands when it was impoverished. O Allah, send blessings and peace and benediction upon Muhammad and upon the Progeny of Muhammad all together, and upon all of the Angels, and upon all of the Prophets, Messengers, martyrs and truthful ones, and upon all of your righteous servants.

Acknowledgements

I would like to thank my parents and my entire extended family, especially, my mother, grandmother, grandfather, aunt, for raising me, teaching me, and providing me with sustenance and opportunities to develop. I pray that Allah (God) instils in my heart special love for them and in their hearts special love for me.

Then, I thank all of my teachers, who sacrificed their time and energy to teach me beneficial knowledge and to facilitate my development as a person and as a professional. Among them, I would like to express my special and sincere gratitude to my doctoral supervisor, my teacher, Dr Coralia Cartis. Her kind and understanding attitude towards me made my PhD days brighter. I thank her for her continued advice, support, and the knowledge and experience she has shared with me. May Allah have mercy upon Coralia Cartis and her family. I would also like to thank my (spiritual) teacher Dr. Riyad Nadwi for the advice, knowledge and wisdom. May Allah have mercy on you and your son.

Then, I express my gratitude to all of my friends who made my difficult days easier and joyful days even more joyful. Among them, my special thanks go to The Oxford University Islamic Society and its members for their hospitality. I immensely benefited from the events and activities you organized. Thank you very much, I hope our paths will cross again. Finally, I would like convey my thanks to Estelle Massart for her help to make the final push and my friend Muhammad Riduan for hosting me during the last days of my PhD and just for the good times together. May Allah have mercy on my friends and their families. Aameen.

Abstract

Though ubiquitous in applications, global optimisation problems are generally the most computationally intense due to their solution time growing exponentially with linear increase in their dimensions (this is the well known/so called 'curse of dimensionality'). In this thesis, we show that this scalability — and sometimes even tractability — challenges can be overcome in the global optimization of functions with low effective dimensionality, that are constant along an (unknown) linear subspace and only vary over the effective (complement) subspace. Such functions can often be found in applications, for example, in hyper-parameter optimization for neural networks, heuristic algorithms for combinatorial optimization problems and complex engineering simulations. We consider using random embeddings to reduce the problem dimension and search space, while attempting to preserve optimal values. In particular, we investigate two randomly reduced (sub)problem formulations that aim to solve the corresponding unconstrained and bound-constrained cases of the global optimization problem.

In our REGO formulation for unconstrained global optimization, a Gaussian random, low-dimensional problem with bound constraints is formulated and solved in a reduced space by means of a(ny) global optimization algorithm. We prove novel probabilistic bounds for the success of REGO in solving the original, low effective-dimensionality unconstrained problem, which show its independence of the (potentially large) ambient dimension and its precise dependence on the dimensions of the effective and randomly embedding subspaces. These results significantly improve existing theoretical analyses by providing the exact distribution of a reduced minimizer and its Euclidean norm and by the general assumptions required on the problem. We validate our theoretical findings by extensive numerical testing of REGO with three types of global optimization solvers, illustrating the improved scalability of REGO compared to the full-dimensional application of the respective solvers.

For the bound-constrained global optimization problem with special structure, a reduced subproblem formulation is investigated that solves the original problem over a Gaussian random low-dimensional subspace subject to affine constraints, so as to preserve feasibility with respect to the given domain. Under reasonable assumptions, we show that the probability that the reduced problem is successful in solving the original, full-dimensional problem is positive. Furthermore, in the case when the objective's effective subspace is aligned with the coordinate axes, we provide an asymptotic bound on this success probability that captures its algebraic dependence on the effective and, surprisingly, ambient dimensions. We then propose X-REGO, a generic algorithmic framework that uses multiple random embeddings, solving the above reduced problem repeatedly, approximately and possibly, adaptively. Using the success probability of the reduced subproblems, we prove that X-REGO converges globally, with probability one, and linearly in the number of embeddings, to a neighbourhood of a constrained global minimizer. Our numerical experiments on special structure functions illustrate our theoretical findings and the improved scalability of X-REGO variants when coupled with state-of-the-art global — and even local — optimization solvers for the subproblems.

In an attempt to improve the scalability of generic global optimization problems, that may not possess low effective dimensionality, we propose to extend the use of the random embeddings framework above to this context. For Lipschitz continuous objectives, we develop a novel analysis that estimates the probability of success of the feasible randomly reduced subproblems based on connections to the field of conic integral geometry. To evaluate the quality of our bound, we compare it to the success of uniform sampling, in the asymptotic regime. Finally, again using our success probability bound, we establish that the X-REGO algorithmic framework applied to the generic bound-constrained global optimization problem is convergent with probability one, and linearly in the number of embeddings, to a neighbourhood of a constrained global minimizer.

Contents

List of Figures

Chapter 1

Introduction

Global optimization is the task of determining a most extreme value of a function over a predefined domain. In mathematical terms, given an *objective function* $f(\boldsymbol{x})$ and a *feasible set* $\mathcal{X} \subseteq \mathbb{R}^D$, global optimization[1] solves

$$\begin{aligned} \min \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in \mathcal{X}. \end{aligned} \tag{GOP}$$

A solution $\boldsymbol{x}^*$ to the global optimization problem is called a *global minimizer* if $f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) = f^*$, for all $\boldsymbol{x} \in \mathcal{X}$, and f^* is called the *global minimum* of f over $\mathcal{X}$.

Global optimization is encountered in a wide variety of applications including portfolio management, protein structure prediction, engineering design, object packing, curve fitting and many more. Unlike local optimization, where one is satisfied with minimizers over a neighbourhood, global optimization requires exploration/consideration of the entire feasible domain. This task is commonly associated with large computational costs, which often grow exponentially with the dimension of the problem rendering global optimization of high-dimensional functions an extremely challenging problem. Along with attempts to devise a generic global optimization algorithm capable of handling high-dimensional problems, researchers have also targeted specific classes of functions, which possess some type of structure often encountered in practice. The main subclass of problems that are 'easy' are convex problems (f is a convex function and $\mathcal{X}$ is a convex set); most convex problems of practical importance are solvable in polynomial time and hence tractable [117]. However, real-life problems are often non-convex, having multiple local and global extrema, or they are black-box, so that their convexity or lack of, cannot be ascertained a priori.

---

[1] The term 'optimization' will refer to 'minimization' since every maximization problem can be cast into minimization problem via $\min f = -\max(-f)$.

1.1 Overview of global optimization algorithms

Research in global optimization can be traced back to the late 1950s, to works by Markowitz and Manne [103], Gomory [67] and Dantzig et al. [39, 38] who addressed linear programming problems with variables accepting only integer values. To solve these, a branch and bound strategy (discussed below) was developed by Land and Doig [90] and Little et al. [99]; the strategy was later adapted and successfully applied to solve problems with continuous variables [50, 136]. A similar branching technique for Lipschitz continuous functions (see definition in (1.3)) was proposed by Piyavskii [129] and Shubert [143] in early 1970s laying foundations of the field now known as Lipschitz Optimization. During the same years, sampling techniques also started to be used for global optimization. For example, clustering methods (see below) were first applied by Becker and Lago [14] in 1970; shortly after, Törn [152, 153] suggested to combine clustering methods with local optimization techniques. Another sampling technique for global optimization that has gained great popularity in recent years is Bayesian Optimization [134] — a statistical approach based on Bayes' rule and other probabilistic tools with its history dating back to works by Kushner [89] (1964), Žilinskas [161] (1975), and Močkus et al. [111] (1978). Other notable methods used for global optimization include interval arithmetic [112, 113], simulated annealing [72, 84, 105] and genetic algorithms [73]. In 1975, the field of global optimization solidified its place as a recognizable discipline with the publishing of the influential volumes "Towards Global Optimization" edited by Dixon and Szegö [42], where they collected a number of papers dedicated specifically to global optimization. Since then, the field has experienced much development and, nowadays, there exist a wide variety of techniques to tackle a global optimization problem. Overviews of global optimization methods can be found in Horst and Pardalos [74], Pintér [128], Floudas [54], Locatelli and Schoen [101] and in the survey by Neumaier [118].

Existing global optimization techniques can be broadly categorized into two groups: deterministic and stochastic. A method is called deterministic[2] if it can provide theoretical guarantees that the value of a calculated solution is larger than the value of the true global minimum by at most ϵ for a pre-specified $\epsilon > 0$ (see [98]). Obtaining a global optimality guarantee on the computed solution in deterministic methods requires searching in the entire feasible domain — a task that is linked to significant computational costs. Whereas in some applications attaining guaranteed performance is crucial (see, e.g., [65, 126]), in many other applications obtaining such guarantees is not required or not even possible[3] (see, e.g.,

---

[2] Deterministic may also refer to techniques that are predictable in their behaviour, that is, for fixed initial parameters an algorithm produces exactly the same results every time it is executed. See [118], for a more detailed classification of global optimization methods.

[3] For example, when the explicit formulation of the objective function is not given.

[150]). In the latter applications, one may be more inclined to employ stochastic methods — methods whose computations are affected by random parameters. Stochastic methods compromise on deterministic guarantees in the hope of achieving good performance with limited computational resources. Deterministic theory in stochastic methods is replaced by statistical analysis and probabilistic guarantees. Results of the following nature may be encountered in stochastic methods: confidence bounds on the predicted global minimum, probabilistic bounds on the number of iterations needed to achieve a certain convergence criteria or the fact that the probability of convergence tends to 1 as the number of iterations grow.

Next, we provide brief descriptions of some of the standard deterministic and stochastic methods focusing on the methods that are directly connected to random techniques developed in this thesis or that are indirectly connected and relevant. This includes deterministic methods such as Branch and Bound and Lipschitz Optimization, Random Search (Stochastic) methods such as Pure/Adaptive Random Search and Simulated Annealing, and methods that can have features of both classes such as Bayesian Optimization and Multi-start. We start with an overview of Branch and Bound.

1.1.1 Branch and Bound

Branch and Bound (B&B) is a general strategy employed to tackle challenging optimization problems. As the name suggests, the B&B strategy has two main components: branching and bounding. At the branching step, B&B splits the problem into smaller sub-problems; the splitting is done in a such way that the resulting sub-problems are 'easier' to solve than the original problem. For example, (GOP) could be split into 2 optimization problems (P_1) and (P_2) each defined over smaller domains $\mathcal{X}_1$ and $\mathcal{X}_2$ such that $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$. Once splitting is completed, B&B performs its second procedure: bounding, which involves obtaining upper and lower bounds[4] for the global minima of sub-problems. The end-goal of the bounding step is to exclude sub-problems that cannot contain the global minimum and thereby shrink the overall search space. One of the remaining (active) sub-problems is then chosen and B&B is applied again. This recursive application of B&B carries on until a certain criterion is satisfied, for example, an approximation of the global minimum is found to a pre-specified accuracy or a maximum computational budget is reached. The overall number of sub-problems that B&B examines before locating a(n approximate) global minimizer depends on various parameters of the problem such as required accuracy of the

---

[4] In the integer programming literature, upper and lower bounds are commonly referred to as primal and dual bounds, respectively. An Integer Programming problem is a global optimization problem where variables are constrained to integer values.

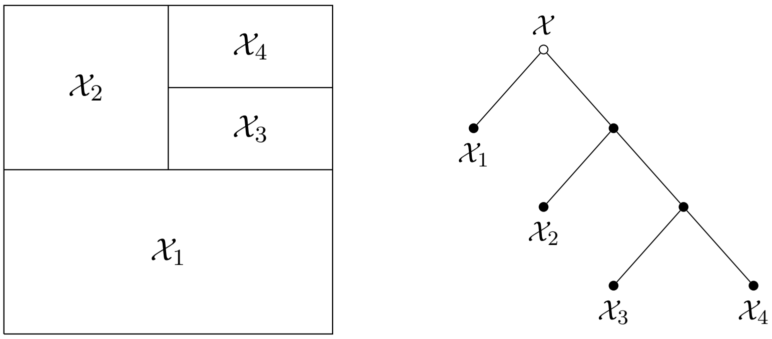


Figure 1.1: An illustration of the B&B branching step. The figure on the left shows the original squared domain $\mathcal{X}$, which has been divided into sets $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{X}_3$ and $\mathcal{X}_4$. The figure on the right shows the tree generated by this particular division.

solution (ϵ in (1.2)), maximal aspect ratio[5] of the boxes and Lipschitz constant (for Lipschitz continuous functions, see (1.3)).

The history of the B&B algorithm in optimization literature starts in the early 1960s [90, 99]. At its earliest stages, B&B was developed and successfully applied for solving discrete optimization problems including integer programming problems with binary variables [11], mixed-integer programming problems[6] [45] and combinatorial optimization problems such as, for example, Travelling Salesman Problem[7] [99] and knapsack problem[8] [88]; see also Mitten [107] and references therein. B&B was later successfully applied to global optimization problems with continuous variables. To make the algorithm suitable for continuous problems, some modifications of the original B&B for discrete problems were required, in particular, a new lower bounding technique for the bounding step had to be developed [50, 136].

We now formally outline the core procedures of the general conceptual B&B algorithm applied to a global optimization problem (GOP) with the global minimum f^*. In describing the B&B steps below, we follow the descriptions in [101, 128]. To help visualize the B&B steps, please refer to Figure 1.1. B&B can be represented by a tree structure, where each parent node corresponds to a (sub-)problem that has been divided into smaller sub-problems represented by the child nodes. The root of the tree corresponds to the original feasible set $\mathcal{X}$.

---

[5] Aspect ratio of a rectangular box is a ratio of the length of its longest side to the length of its shortest side.

[6] Mixed-integer problems are a class of problems which, in addition to continuous variables, have discrete variables.

[7] Travelling Salesman Problem is stated as follows: given a list of cities and the distances between them, find the shortest route that visits each city exactly once and returns to the starting city.

[8] Knapsack problem is motivated by a mundane, practical task: given a list of items each with a certain weight and value, and a bag with a certain weight limit, how many of each item one needs to put in the bag such that the total value of the items in the bag is maximized while the total weight of the bag does not exceed the limit.

Initialization step. Let $k = 1$. Let $\mathcal{X}_1 = \mathcal{X}$ be the initial feasible set. Denote by $\mathcal{A}^1 = \{\mathcal{X}_1\}$ the set of active subsets. For $k \geq 1$, execute the following steps.

Bounding step. For each $\mathcal{X}_i \in \mathcal{A}^k$, compute an upper bound $ub(\mathcal{X}_i)$ and a lower bound $lb(\mathcal{X}_i)$ for the global minimum of the following problem:

$$\begin{aligned} \min \;\; & f(\boldsymbol{x}) \\ \text{subject to} \;\; & \boldsymbol{x} \in \mathcal{X}_i. \end{aligned}$$

After obtaining estimates for each active subset $\mathcal{X}_i$, we can define upper and lower bound estimates for f^*:

$$lb^k := \min_{\mathcal{X}_i \in \mathcal{A}^k} lb(\mathcal{X}_i) \leq f^* \leq \min_{\mathcal{X}_i \in \mathcal{A}^k} ub(\mathcal{X}_i) := ub^k. \tag{1.1}$$

Termination step. After determining the values of lb^k and ub^k, we can add the following stopping criteria:

$$ub^k - lb^k \leq \epsilon \tag{1.2}$$

for a pre-specified tolerance $\epsilon > 0$. In the case of such termination, the value of ub^k is ϵ-optimal in the sense that $ub^k - f^* \leq \epsilon$, which must be satisfied due to (1.1).

Fathoming step. The term fathoming in B&B refers to the process of removing subsets from $\mathcal{A}^k$ that cannot contain the global minimum. For instance, a set $\mathcal{X}_i$ can be fathomed if $lb(\mathcal{X}_i) \geq ub^k$. Let $\mathcal{R}^k$ denote the collection of subsets that are to be fathomed. Update $\mathcal{A}^k := \mathcal{A}^k \setminus \mathcal{R}^k$.

Selection step. Select one or several subsets from $\mathcal{A}^k$ and denote this collection by $\mathcal{X}^k$. Often, the subset that has the lowest $lb(\mathcal{X}_i)$ from the remaining subsets in $\mathcal{A}^k$ is chosen [101, 128]. Such a choice is governed by a rationale that the subsets with small lower bound estimates might contain good solutions and therefore its branching might produce significant improvements on ub^k and subsequent fathoming of greater number of subsets.

Branching step. Partition each $\mathcal{X}_i$ from $\mathcal{X}^k$ into a finite number of subsets. Add the new list of subsets to $\mathcal{A}^k$. Return to Bounding step and update $k := k + 1$.

The bounding step is perhaps the most crucial step of the B&B strategy. Obtaining good upper and lower bound estimates is key to achieving rapid fathoming of sub-problems and, thus, faster convergence. Obtaining an upper bound is conceptually easy: a value of f at any feasible point gives an upper bound. We note here that for certain domains (for example, defined by non-linear constraints) finding a feasible point is a challenging task in itself [118]. It can be tackled by formulating an optimization problem with zero objective and constraints given by the domain in which a feasible point is sought. This optimization problem, however, could be as hard as a global optimization problem [118]. Thus, in B&B, division into simple (sub-)regions is preferable, for example, rectangles or balls for which finding a feasible point is a trivial task. A lower bound for a sub-problem is usually found by formulating an 'easy' optimization problem whose global minimum underestimates the global minimum of the sub-problem. These underestimating functions are typically given by convex functions, which can be efficiently solved by local optimization methods [23, 117]. Lower bound estimates can also be found by exploiting a known structure of the problem. This is done, for example, in Differential Convex (D.C.) Programming (see [75]) — an optimization problem with objective and constraint functions that can be written as a difference of two convex functions. A lower bound in D.C. Programming can be obtained, for example, by solving a linear programming problem (see [75]).

Later in the numerical parts of the thesis, we use a B&B type solver called BARON (Branch And Reduce Optimization Navigator) [139, 148]. BARON can handle problems with both integer and continuous variables, with linear and non-linear constraints. Prior to implementing the main B&B procedure, BARON allows the user to run a pre-processing step in which it executes a multi-start local optimization heuristic (see Section 1.1.4) to produce good feasible solutions. These solutions can then be used as upper bound estimates in the main B&B procedure. For more description of BARON, see page 51 in Section 3.3. Other global optimization solvers that employ B&B strategy include α-BB method [2, 7], ANTIGONE [106], GlobSol [82] and LINGO [61].

1.1.2 Lipschitz Optimization

Lipschitz Optimization refers to a class of optimization methods characterized by two key features:

- Lipschitz optimization methods are sampling techniques designed to optimize black-box functions, i.e., functions that do not have explicit formulations and only $f(\boldsymbol{x})$ is observed when $\boldsymbol{x}$ is sampled. First- and higher-order derivatives are not available.

- Lipschitz optimization methods operate under an assumption that the objective function f is Lipschitz continuous, that is, there exists a (Lipschitz) constant $L > 0$ such that
$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2 \text{ for all } \boldsymbol{x} \text{ and } \boldsymbol{y} \text{ in } \mathcal{X}. \tag{1.3}$$

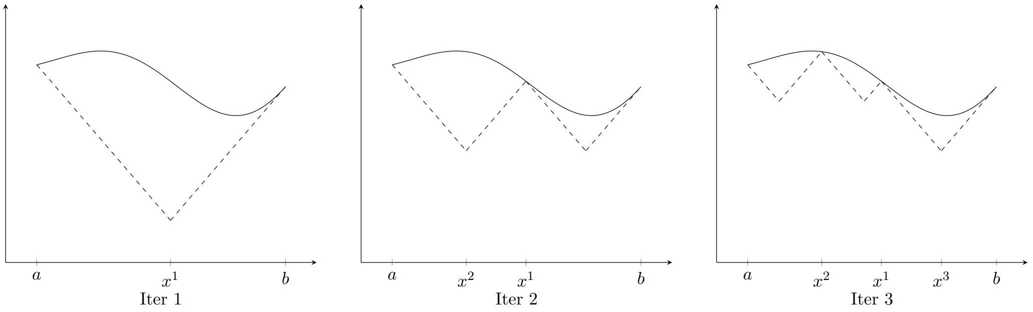


Figure 1.2: An illustration of Piyavskii and Shubert's algorithm.

> The distinctive feature of these methods is that their algorithmic procedures heavily rely on (1.3). In particular, the bounds that are implied by (1.3) are used to decide on subsequent sampling points and when to terminate the algorithms.

We start the description by mentioning key implications of (1.3). Note that the bound in (1.3) can be used to lower bound the global minimum f^*. If $r := \max_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{X}} \|\boldsymbol{x} - \boldsymbol{y}\|$ the diameter of the domain $\mathcal{X}$ is known, then, (1.3) yields

$$f^* \geq f(\boldsymbol{x}) - Lr \text{ for all } \boldsymbol{x} \in \mathcal{X}.$$

This observation gives rise to a powerful principle: the approximation of f^* is more accurate for smaller search spaces. This principle lies at the core of Lipschitzian optimization methods which divide the original search space $\mathcal{X}$ into smaller spaces $\mathcal{X}_i$, $i = 1, 2, \ldots, N$ such that

$$\mathcal{X} = \bigcup_{i=1}^{N} \mathcal{X}_i$$

and determine lower bounds F_i of f on each $\mathcal{X}_i$. The lowest lower bound $F = \min_{1\leq i\leq N} F_i$ is considered as the lower bound on f over the entire $\mathcal{X}$. As the division gets finer, the difference between F and f^* tends to zero (see [129]).

Based on this principle, Piyavskii [129] and Shubert [143] developed a Lipschitzian optimization algorithm for one-dimensional functions, which we now describe closely following the description in [78]. Let us consider a global optimization problem of one-dimensional function f defined over the interval $[a, b]$. We can lower bound f on $[a, b]$ using (1.3). Let x be any point in $[a, b]$. By applying (1.3) for points x and a and points x and b, we obtain two inequalities:

$$\begin{aligned} f(x) &\geq f(a) - L(x - a), \\ f(x) &\geq f(b) + L(x - b). \end{aligned} \tag{1.4}$$

The right-hand side of these two inequalities are two lines with slopes $-L$ and L (see Figure 1.2, Iter 1). Together these lines form a V-shaped linearly-piecewise function that lies below f and that intersects f at the endpoints of the interval a and b. The two lines meet at the bottom of the V at a point $(X(a,b), F(a,b))$, where

$$X(a,b) = (a+b)/2 + (f(a) - f(b))/(2L),$$
$$F(a,b) = (f(a) + f(b))/2 - L(b-a)/2.$$

The value of $F(a,b)$ is considered as the lower bound of f. The method further divides the interval $[a,b]$ into two intervals $[a,x^1]$ and $[x^1,b]$, where $x^1 = X(a,b)$. New V-shaped underestimates for f are formed over these two intervals based on (1.4) and the corresponding values of X and F are calculated. The method proceeds further by choosing the interval with the lowest value of F. In this case, F's of both intervals are equal, in which case we choose any of the two intervals, for example, $[a,x^1]$ and split it at $x^2 = X(a,x^1)$; see Iter 2 in Figure 1.2. Now, the domain is divided into three intervals $[a,x^2]$, $[x^2,x^1]$ and $[x^1,b]$. Among these three intervals, the interval $[x^1,b]$ possesses the lowest value of F and thus the next point is chosen at $x^3 = X(x^1,b)$. At every subsequent iteration, the algorithm samples the point that corresponds to the minimum of the piecewise linear approximation. As the number of samples increases, the linear approximation becomes more accurate. The algorithm terminates when the difference between the lowest value of the approximation and the current best sample is less than a specified tolerance.

The algorithm can be extended to multivariate objective functions defined over rectangular regions as was done, for example, in Pintér [127] and Galperin [59]. The initialization step of their approaches, however, requires sampling 2^D vertices of the search space, which leads to a fast rising number of samples as D increases. A second major issue associated with classical Lipschitzian algorithms is the need to specify the Lipschitz constant L in the algorithm, which for many objective functions encountered in real-life applications may not be available or not easily computable. To address these issues, Jones et al. [78] developed an algorithm called DIRECT (DIviding RECTangles) — a modification of the Piyavskii and Shubert's algorithm. Jones et al. propose a new partition method which requires evaluation of only one point (midpoint) in the search space while still being able to provide lower bounds on f using (1.4). To overcome the issue of specifying L, DIRECT "carries out simultaneous searches using all possible constants from zero to infinity" [78]. For more details, see [78]. It is important to mention that although DIRECT avoids sampling 2^D points in the search domain to initialize the search, it does not mean that it achieves good performance in high-dimensional spaces and convergence to the global minimum may still be slow [78]. We use DIRECT in our numerical experiments later on in the thesis. We provide additional description of the solver on page 51 in Section 3.3.

1.1.3 Bayesian Optimization

Bayesian Optimization (BO) is another class of optimization methods designed to tackle black-box functions. BO is a probabilistic model-based strategy particularly suited for problems where only a small number of function evaluations is allowed due to high costs[9] associated with each evaluation. BO is a sequential process that involves two main procedures. First, it constructs an approximation surrogate function for the objective function f based on the previously observed data; typically, when f is continuous it uses a Gaussian process. Then, relying on the constructed surrogate function, it determines the next sampling point by optimizing a so-called acquisition function. It iterates between these two steps until a certain limit on the number of function evaluations is reached.

The term Bayesian Optimization was coined by Močkus [109] (1982) but the history of the field begins earlier with the works of Kushner [89] (1964), Žilinskas [161] (1975) and Močkus et al. [111] (1978), who noted the usefulness of Bayesian methods in global optimization. BO has since been used in fields such as engineering system design [110], drug discovery [115], robotics [100], animation design [25], and reinforcement learning [24]. The field has received significant attention after the popular work of Jones et al. [79] in 1998 on the Efficient Global Optimization algorithm, and more recently, within the machine learning community after the work of Snoek et al. [144] (2012) who noted that BO can be used to train neural networks.

In what follows, we describe the Bayesian Optimization steps in more detail. We start with the description of a Gaussian process. The term Gaussian process refers to a belief that f is a random sample from a pool of functions that together form a normal distribution. More precisely, given the matrix $\boldsymbol{X}^N = (\boldsymbol{x}^1 \ \boldsymbol{x}^2 \ \ldots \boldsymbol{x}^N)$ of N observed points $\boldsymbol{x}^i \in \mathbb{R}^D$ and the vector $\boldsymbol{f}^N = (f(\boldsymbol{x}^1) \ f(\boldsymbol{x}^2) \ldots f(\boldsymbol{x}^N))^T$ of the corresponding values of f, Gaussian process assumes that $f(\boldsymbol{x}^1), f(\boldsymbol{x}^2), \ldots, f(\boldsymbol{x}^N)$ follow a multivariate normal distribution:

$$\boldsymbol{f}^N \sim \mathcal{N}(\boldsymbol{m}(\boldsymbol{X}^N), \boldsymbol{K}(\boldsymbol{X}^N, \boldsymbol{X}^N)), \tag{1.5}$$

Here, $\boldsymbol{m}(\boldsymbol{X}^N)$ denotes the vector $(m(\boldsymbol{x}^1) \ m(\boldsymbol{x}^2) \cdots m(\boldsymbol{x}^N))^T$ formed by evaluating a prespecified mean function $m : \mathbb{R}^D \to \mathbb{R}$ at the observed points and $\boldsymbol{K}(\boldsymbol{X}^N, \boldsymbol{X}^N)$ is a covariance matrix whose (i, j)th entry is given by $k(\boldsymbol{x}^i, \boldsymbol{x}^j)$ — a user-specified kernel function $k : \mathbb{R}^{D\times D} \to \mathbb{R}$ evaluated for points $\boldsymbol{x}^i$ and $\boldsymbol{x}^j$. For a kernel k, the squared exponential function is typically used:

$$k(\boldsymbol{x}^i, \boldsymbol{x}^j) = \exp\left(-\frac{1}{2}\|\boldsymbol{x}^i - \boldsymbol{x}^j\|_2^2\right). \tag{1.6}$$

---

[9] In some applications, for example, it may take several hours to evaluate a function or there is monetary cost attached to each evaluation; see Frazier [56].

Note that the squared exponential function k is larger for points closer to each other; this imposes a belief that function values of points that are closer must have greater correlation. Furthermore, with such k, $\boldsymbol{K}$ is a positive semi-definite matrix for any collection of points (see [134]). For the mean function m, a constant function is commonly chosen, for example, $m \equiv 0$.

In Bayesian statistics, the relation in (1.5) is referred to as the *prior distribution*. We now would like to determine the conditional distribution of $f(\boldsymbol{x}^*)$ at a new point $\boldsymbol{x}^*$ that we are about to sample conditioned on the observed $\boldsymbol{f}^N$. Note that at this stage we treat $\boldsymbol{x}^*$ as a variable; the goal here is to find a closed-form expression for the distribution of $f(\boldsymbol{x}^*)$ for all possible points $\boldsymbol{x}^*$. To derive the conditional distribution, we first form a prior distribution of $(\boldsymbol{f}^N \;\; f(\boldsymbol{x}^*))^T$ using (1.5). Then, using Bayes' rule, for the conditional distribution $f(\boldsymbol{x}^*)|\boldsymbol{X}^N, \boldsymbol{f}^N$, we obtain (see [134, Chapter 2.1] for details)

$$\begin{aligned} f(\boldsymbol{x}^*)|\boldsymbol{X}^N, \boldsymbol{f}^N &\sim \mathcal{N}(\mu(\boldsymbol{x}^*), \sigma(\boldsymbol{x}^*)), \\ \mu(\boldsymbol{x}^*) &= \boldsymbol{k}(\boldsymbol{x}^*, \boldsymbol{X}^N)\boldsymbol{K}(\boldsymbol{X}^N, \boldsymbol{X}^N)^{-1}(\boldsymbol{f}^N - \boldsymbol{m}(\boldsymbol{X}^N)) + m(\boldsymbol{x}^*), \\ \sigma(\boldsymbol{x}^*) &= k(\boldsymbol{x}^*, \boldsymbol{x}^*) - \boldsymbol{k}(\boldsymbol{x}^*, \boldsymbol{X}^N)\boldsymbol{K}(\boldsymbol{X}^N, \boldsymbol{X}^N)^{-1}\boldsymbol{k}(\boldsymbol{X}^N, \boldsymbol{x}^*), \end{aligned} \tag{1.7}$$

where $\boldsymbol{k}(\boldsymbol{x}^*, \boldsymbol{X}^N) = \boldsymbol{k}(\boldsymbol{X}^N, \boldsymbol{x}^*)^T = (k(\boldsymbol{x}^*, \boldsymbol{x}^1) \;\; k(\boldsymbol{x}^*, \boldsymbol{x}^2) \;\; \ldots \;\; k(\boldsymbol{x}^*, \boldsymbol{x}^N))$. The relation in (1.7) is called the *posterior distribution*.

Once the posterior distribution has been computed, BO begins its second phase: determining the next query point. The next point is chosen based on the acquisition function. Many different choices for the acquisition function exist, for example, the expected improvement [108], Thompson sampling [151], probability of improvement [77] and upper confidence bounds [146]. Here we present results for the most commonly used acquisition function, which is the *expected improvement*. The expected improvement is defined as

$$\mathrm{EI}^N(\boldsymbol{x}^*) := \mathbb{E}[\max\{0, f_{min}^N - f(\boldsymbol{x}^*)\}|\boldsymbol{X}^N, \boldsymbol{f}^N], \tag{1.8}$$

where $f_{min}^N = \min_{i \le N} f(\boldsymbol{x}^i)$ and where $f(\boldsymbol{x}^*)$ conditioned on $\boldsymbol{X}^N$ and $\boldsymbol{f}^N$ has the distribution (1.7). The next sampling point is chosen as the point at which $\mathrm{EI}^N(\boldsymbol{x}^*)$ achieves its maximum value:

$$\boldsymbol{x}^{N+1} = \arg\max \mathrm{EI}^N(\boldsymbol{x}^*). \tag{1.9}$$

The value of the expected improvement is larger for points in poorly explored regions or points with high expected value. Thus, point $\boldsymbol{x}^{N+1}$ favours under-explored regions, where a potential global minimizer may be located, and/or the vicinity of the current best solution, where possibly an even better solution is hidden. This property of the expected improvement is illustrated in Figure 1.3, where the sixth query point has been chosen in the area with a large gap between nearest sampled points.

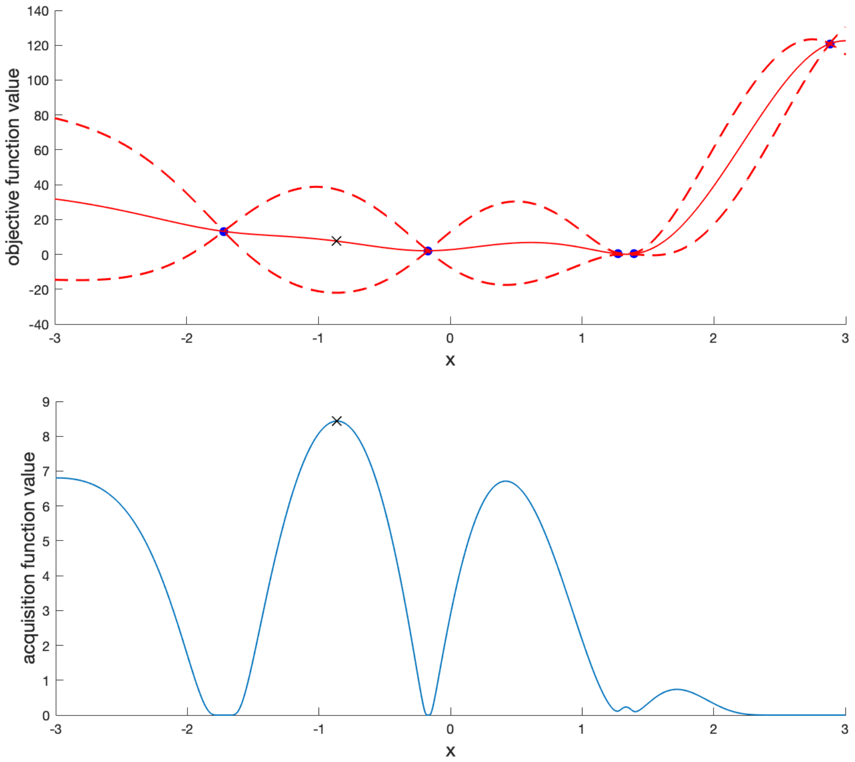


Figure 1.3: An illustration of a BO step. The top figure shows five sampled points in blue and the mean of the posterior distribution in solid red with confidence regions in dashed red. The bottom figure depicts the acquisition function, which attains its maximum at a point denoted by ×; the x-component of this point is the next sampling point.

The expected improvement (1.8) has a closed-form expression (see, e.g., [79]). Evaluations of the expected improvement are inexpensive and its first and second derivatives are easily available. Hence, the optimization problem (1.9) does not pose a challenge and can be solved by off-the-shelf optimization methods such as, for example, DIRECT [78], though it may still be computationally consuming.

1.1.4 Multi-start methods

Multi-start methods are a class of techniques that tackle a global optimization problem (GOP) by applying the same set of rules each time with different initialization parameters. A multi-start method, in its basic form, consists of two main procedures: generation of a feasible solution and improvement of the solution through an application of a fast and efficient algorithmic procedure. This improvement procedure is commonly a local optimization framework in which the generated solution takes the role of a starting point. Such a multi-start method would apply a local optimization algorithm from many different starting points which are typically initialized in such a way as to find as many local minima of the problem as possible. The hope is that one of the found local minima found will be a global one. Let us now review key state-of-the-art techniques for local optimization that are often used in multi-start methods.

Local optimization. Local optimization algorithms are widely used within multi-start as well as other global optimization methods. Different local optimization methods have been developed to tackle different classes of objectives (see [36, 123]). Existing methods can be classified into two broad groups: derivative-based and derivative-free. Derivative-based local optimization methods, as the name suggests, use available derivative information about the objective function f to drive an initial starting point towards a local minimizer. They perform iterative procedures, which at each iteration, choose a point that is 'better' than the current one, where, 'better' can mean, for example, that the function value of a new point is smaller than the function value of the current one. The choice of a new point is typically found by minimizing a linear or quadratic model of f around the current point. Linear models are built based on the gradient of the objective f, whereas quadratic models also utilize information about the Hessian[10] of f.

In line-search methods, for instance, these models are then used to determine a suitable search direction $\boldsymbol{s}$ along which they move from the current point, say $\boldsymbol{x}$, to a new point $\boldsymbol{x}'$. More precisely, $\boldsymbol{x}' = \boldsymbol{x} + \alpha \boldsymbol{s}$, where $\alpha > 0$ can be assigned a value that corresponds to the minimum of $f(\boldsymbol{x} + \alpha \boldsymbol{s})$ (as in *exact line-search*) or determined through an iterative process as in backtracking Armijo procedure. Popular line-search methods include steepest-descent, *Newton* and *Quasi-Newton* methods (see [123] for details). Another approach for choosing $\boldsymbol{x}'$ is based on a model trusting principle — a class of methods known as *Trust-Region* (TR) [35]. At each iteration, a typical TR method assigns a quadratic model of a certain radius around the current point $\boldsymbol{x}$ and if certain progress conditions are met, the radius is increased and $\boldsymbol{x}'$ is chosen as a point which the model predicts would give the best function decrease. If such conditions are not satisfied then the radius is decreased and the procedure is repeated.

If a local optimization problem involves constraints, additional techniques must be implemented to ensure feasibility of computed solutions. Constraints in local optimization can be handled, for example, by *Penalty* or *Augmented Lagrangian methods* (see [123, Chapter 17]), which formulate a new unconstrained optimization problem that incorporates the constraints into the objective function. If constraints form a simple set, one can also use *gradient projection methods* [83, Chapter 5], which ensure feasibility by projecting every (unconstrained) optimization step onto the feasible set.

---

[10] The Hessian of f is a matrix whose (i, j)th entry is given by $\partial^2 f / \partial x_i \partial x_j$.

State-of-the-art derivative-based local optimization solvers include KNITRO [28], Ipopt [162], SNOPT [63, 64] and GALAHAD [69]. We use KNITRO later on in the thesis to conduct our numerical experiments. For a description of KNITRO, see page 51, Section 3.3.

For some real-life problems such as tuning of algorithmic parameters [8], engineering design problems [21, 22], dynamic pricing [94] and more (see [36]), derivative-based local optimization methods cannot be employed due to no — or lack of access to— derivative information. To tackle such problems, derivative-free optimization (DFO) methods have been developed. The underlying algorithmic principle of DFO methods is similar to the one of derivative-based: iteratively go downhill, but the difference is that they try do so using only function evaluations. Some DFO methods mimic classical methods in that they construct linear or quadratic models around the current point. *Implicit filtering* methods [83], for example, build these models by approximating the derivatives in classical line-search methods using finite differencing [27, Chapter 4]. DFO trust-region methods replace the quadratic model formed based on the gradient and Hessian with a quadratic model built by interpolation of function values of a number of previously sampled points. 'Non-model'-based approaches to generate search directions also exist. In *pattern* or *direct search* methods [87], for example, to generate a new point, a set of so-called poll directions is explored; this set is built in such a way as to ensure that at least one of the poll directions is a descent direction. For more information on DFO methods, consult [36, 87, 116, 135].

Let us now resume our discussions on multi-start methods, which — as we mentioned above — typically comprise of two procedures, namely, initialization of starting points and application of local optimization procedures from these points. Since the success of a multi-start method chiefly depends on the choice of the starting points, much of the focus in the literature has been on developing efficient initialization strategies for these. In general, the starting points can be initialized in one of three ways: deterministically (for example, chosen from a grid), randomly (for example, sampled uniformly in $\mathcal{X}$) or adaptively, i.e. chosen based on the information from previous local runs. The advantage of deterministic and random initialization strategies is that they allow local optimization procedures to be run in parallel leading to reduction in the wall-clock running time[11]. The main disadvantage of these initialization strategies, however, is that many starting points may end up converging to the same local minimizer, thus wasting much of the computational resources. To address this issue, adaptive ways to choose starting points have been proposed such as tunnelling [95, 96, 160, 170] and clustering [14, 80, 154] methods. The aim of both methods is to create a sequence of starting points such that an application of local minimization procedures with

---

[11] True elapsed real time between the start and end of the process.

these points will lead to different local minima. In what follows, we describe these two methods in more detail.

Tunnelling method. The basic idea of the tunnelling method is to create a sequence of feasible starting points $\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^k$, which after an application of local optimization procedures, produce a sequence of local minimizers $\boldsymbol{x}_*^1, \boldsymbol{x}_*^2, \ldots, \boldsymbol{x}_*^k$ such that each subsequent minimizer is an improvement on the previous, that is, $f(\boldsymbol{x}_*^1) > f(\boldsymbol{x}_*^2) > \cdots > f(\boldsymbol{x}_*^k)$. Such sequence of starting points is generated by finding a point $\boldsymbol{x}^k \in \mathcal{X}$, at each iteration k, that satisfies $\boldsymbol{x}^k \neq \boldsymbol{x}_*^{k-1}$ and $f(\boldsymbol{x}^k) = f(\boldsymbol{x}_*^{k-1})$. In other words, a new starting point that is no worse than the previous local minimizer is sought. The method searches for such a point by creating an auxiliary (or tunnelling) function defined as

$$T^k(\boldsymbol{x}) = \frac{f(\boldsymbol{x}) - f(\boldsymbol{x}_*^{k-1})}{\|\boldsymbol{x} - \boldsymbol{x}_*^{k-1}\|^{\alpha^k}}$$

for some positive parameter α^k. Then, the equation

$$T^k(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \mathcal{X} \tag{1.10}$$

is solved and a solution of (1.10) is assigned as the new starting point. Note that, due to the definition of $T^k(\boldsymbol{x})$, the current best minimizer $\boldsymbol{x}_*^{k-1}$ cannot be a solution of (1.10). Local minimizers $\{\boldsymbol{x}_*^i\}_{i<k}$ of the previous iterations are also eliminated because $f(\boldsymbol{x}_*^i) - f(\boldsymbol{x}_*^k) > 0$ for $i < k$. The tunnelling method terminates when (1.10) has no solution, which would imply that $\boldsymbol{x}_*^k$ is the global minimum.

Despite these nice improvement properties of the tunnelling method, solving (1.10) or verifying that it does not have any solutions could be as difficult as determining the global minimum of the original global optimization problem (GOP) itself (indicating that equation (1.10) is essentially a reformulation of (GOP)). Note also that choosing proper values for parameters α^k is another issue of the method that has not been well-understood. Furthermore, for a general problem, no solid convergence theory has been derived and no well-established usable stopping rule exists. "Thus, lacking any sort of guarantee, the method is at best of some heuristic value" [80, p. 649].

Clustering methods. Just like the tunnelling method, clustering methods aim to generate a sequence of different local minimizers. The underlying principle of clustering methods is related to the notion of basin of attraction. In layman's terms, a basin of attraction is defined as a subset of the feasible domain $\mathcal{X}$ that consists of all points whose descent path lead to the same local minimizer. Clustering methods are founded on the idea that, given a basin of attraction, (in a theoretical sense) one does not need to apply a local descent algorithm from more than one point in the basin to find its local minimum. Different variants of the clustering algorithm have been proposed (see [10, 80]), but the general clustering algorithm roughly follows the below procedure:

1. *Initialization step.* Let $S = \varnothing$. For $k \geq 1$, execute the following steps.
2. *Generation step.* Sample N points uniformly at random in $\mathcal{X}$. Add these points to S.
3. *Concentration step.* Retain only a fraction of points in S with the lowest function values (see [14]). Such selection will produce groups of points concentrated around local minima. Concentration can also be performed through running a few decent steps from each sampled point to move them closer to local minima (see [154]).
4. *Clustering step.* From each group of points formed in the previous step, select a point that has the lowest function value. Let X denote the collection of the selected points and let $\boldsymbol{x} \in X$. For each $\boldsymbol{y}$ in S, check if

   $$\|\boldsymbol{x} - \boldsymbol{y}\| \leq \delta \text{ for some } \delta > 0. \tag{1.11}$$

   If (1.11) holds, then $\boldsymbol{y}$ is assigned to the same cluster with $\boldsymbol{x}$. This procedure is repeated for all $\boldsymbol{x} \in X$ that have not been clustered. After clusters have been formed, from each cluster choose one point with the lowest function value and collect all these points in the set C.
5. *Optimization step.* Run a local optimization algorithm starting from each point in C. New local minima found are added to S.
6. *Termination step.* If a termination criterion is not satisfied, return to Generation step and sample uniformly at random or in an informed way utilizing information from the previous iteration (e.g., keeping old points and sampling new points in under-explored regions).

As any other general global optimization method, clustering methods suffer from the ‘curse of dimensionality’: an exponential number in dimension D samples is required to be able to form meaningful clusters. Recent work by Bagattini et al. [9, 10], following trends in machine learning, addressed the curse of dimensionality by forming clusters in smaller d-dimensional feature space, which they assume characterizes the full dimensional problem well. They successfully apply their method on the atomic cluster structure prediction problem in computational chemistry and sphere packaging problem in geometry.

1.1.5 Random Search methods

Multi-start methods belong to the class of Random Search (RS) if the starting points are initialized at random. Other algorithms in this class include the following.

Pure Random Search. Pure Random Search (PRS) [26, 42] is perhaps the most basic RS method developed for global optimization. PRS repeatedly samples random points in $\mathcal{X}$, typically from the uniform distribution, accepting a new point only if it is an improvement on the current best one. Given that points are sampled independently, one can easily prove that PRS produces a sequence of points that converges to a solution — that is within ϵ distance from the global minimum — with probability one (see [145, 172]). However, the convergence is generally very slow: assuming that the volume of $\mathcal{X}$ is equal to 1, the expected number of samples needed to converge to a ball of radius ϵ around a global minimizer is proportional to $(1/\epsilon)^D$.

Pure Adaptive Search. Pure Adaptive Search (PAS) is a modification of PRS. It was initially used for convex problems in [125], but was later applied and analysed for global optimization problems with Lipschitz continuous objective functions [173]. Just like PRS, PAS generates a sequence of points $\boldsymbol{x}^1, \boldsymbol{x}^2, \dots \in \mathcal{X}$, but instead of sampling in the entire domain $\mathcal{X}$ at every iteration $k \geq 1$, PAS samples uniformly in the level set $LS(\boldsymbol{x}^k) = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) < f(\boldsymbol{x}^k)\}$. Such algorithmic procedure ensures that every subsequent sampled point is an improvement on the previous one. Whereas in PRS sampling uniformly in $\mathcal{X}$ could be very easy (e.g., $\mathcal{X}$ is a rectangle), sampling uniformly in a level set $LS(\boldsymbol{x}^k)$ is in general very difficult [125]. However, the theoretical analysis of PAS demonstrates that the ability of sampling in improving regions bears significant advantages. For example, for Lipschitz continuous functions, the expected number of samples needed to converge to a point that is within ϵ distance from the global minimum is linear in D [172, Theorem 2.9].

Simulated Annealing. Simulated Annealing is a probabilistic sampling technique that is motivated by a thermodynamic process in condensed matter physics called annealing — a procedure where a solid in a heat bath is heated to a high temperature and then slowly cooled. Such process allows the particles in the solid to arrange themselves in a particular lattice structure that corresponds to the lowest energy ground state. Simulated Annealing algorithm is based on the Metropolis procedure introduced in [105] as a technique to simulate the behaviour of particles at a fixed temperature. The Metropolis procedure with decreasing temperatures can then be used to achieve the low energy ground state (see [84]). See Kirkpatrick et al. [84], for the description of similarities between the fields of optimization and condensed matter physics motivating the use of the Metropolis procedure in optimization.

Simulated Annealing algorithm generates a sequence of points $\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots$ according to the following rule:

$$\boldsymbol{x}^{k+1} = \begin{cases} \boldsymbol{z}^{k+1} & \text{with probability } \min\left\{1, \exp\left(\frac{f(\boldsymbol{x}^k) - f(\boldsymbol{z}^{k+1})}{T^k}\right)\right\} \\ \boldsymbol{x}^k & \text{otherwise,} \end{cases}$$

where $\boldsymbol{z}^{k+1}$ is a random point whose distribution depends on $\boldsymbol{x}^k$ and, possibly, also on T^k, and where T^k is the temperature parameter, which progressively decreases as k grows. Note that if $f(\boldsymbol{z}^{k+1}) < f(\boldsymbol{x}^{k+1})$, i.e. $\boldsymbol{z}^{k+1}$ is a better point, $\boldsymbol{z}^{k+1}$ is accepted with probability one. If, however, $f(\boldsymbol{z}^{k+1}) > f(\boldsymbol{x}^{k+1})$ the procedure accepts $\boldsymbol{z}^{k+1}$ with probability $\exp\big(f(\boldsymbol{x}^k) - f(\boldsymbol{z}^{k+1})/T^k\big)$. For a fixed temperature parameter, this acceptance probability is smaller for larger difference between $f(\boldsymbol{z}^{k+1})$ and $f(\boldsymbol{x}^k)$. By occasionally accepting worse points, Simulated Annealing algorithm allows a wider search of the global minimum solution. It was shown that the algorithm converges to the global minimum with probability one provided a suitable decrease rate of T^k to 0 (see [157, 171]). If T^k is decreased too fast, the procedure is likely to get trapped in local minimum [171]. See [101, 171, 157] for more details.

For more details on RS methods for global optimization, see [101, Chapter 3].

Random search methods in local optimization. RS methods in local optimization are commonly used to solve large-scale problems for which conventional deterministic methods fail to produce suitable solutions. These large-scale optimization problems are commonly encountered, for example, in machine learning where problems can have millions to billions of variables. Depending on the type of a problem being considered, various different methods have been developed. Many of these methods are based on one common principle, that is, reducing the number of variables (or other components of the problem) before applying optimization procedures. Below we list a few known methods.

1. *Block Coordinate Descent.* These methods tackle a problem with a large number of variables by optimizing only a small number (block) of variables at a time. First, variables are divided into groups and then a typical Block Coordinate Descent procedure alternates between the following two steps: i) choose a group (e.g. randomly [53] or in a deterministic (cyclic) manner [169]) ii) optimize over the variables in the chosen group. For more information, consult [13, 168] and the references therein.
2. *Stochastic Gradient Descent.* These methods are employed to optimize functions given as a sum of a large number of simple functions. The idea of Stochastic Gradient

Descent, as the name suggests, is based on the conventional gradient descent methods. However, instead of computing a full gradient — which requires computation of gradients of each function in the sum at each iteration — Stochastic Gradient Descent computes the gradient of only one, chosen randomly. See [68, 102] and the references therein.

3. *Sketching.* Sketching is a popular method for tackling the large scale least-squares regression problem (see [123, Chapter 10]) — a problem of finding a solution to an overdetermined linear system (with a very large number of rows) that minimizes the Euclidean norm of the residual. Sketching technique premultiplies the linear system by a random fat matrix that has a small number of rows to reduce the dimension of the system and make it tractable for subsequent optimization procedures. For more details see [133, 31, 167].

For more on RS methods for local optimization, please refer to [91, 101, 135].

In Chapter 6 of this thesis, we propose an RS algorithm for global optimization that operates on a similar principle as a typical local RS method. To find a solution to the global optimization problem, our algorithm alternates between two steps: i) Reduction of the dimension of the problem by embedding a random subspace ii) Optimization of the objective function on the subspace locally or globally.

1.2 Functions with low effective dimensionality - an introduction

In this thesis, we address both unconstrained ($\mathcal{X} = \mathbb{R}^D$) and bound-constrained[12] ($\mathcal{X} = [-1, 1]^D$) formulations of (GOP). Namely, we focus on the global optimization problems

$$f_U^* = \min_{\boldsymbol{x} \in \mathbb{R}^D} \; f(\boldsymbol{x}) \qquad \text{(UP)} \qquad\qquad \begin{aligned} f^* = \;& \min \; f(\boldsymbol{x}) \\ & \text{s. t. } \; \boldsymbol{x} \in \mathcal{X}, \end{aligned} \qquad \text{(P)}$$

where $f : \mathbb{R}^D \to \mathbb{R}$ is continuous, deterministic and possibly non-convex. For (UP), we additionally assume that there exists $\boldsymbol{x}^* \in \mathbb{R}^D$ such that $\min_{\boldsymbol{x} \in \mathbb{R}^D} f(\boldsymbol{x}) = f(\boldsymbol{x}^*) = f_U^*$. This implies that f is bounded below, namely, $f_U^* > -\infty$, and that the minimum in (UP) is attained (not all minimizers are at infinity). Note that, for (P), this property is implied by the Weierstrass extreme value theorem (see, e.g., [130, Chapter 3]), which says that if f is a continuous function defined over a compact set $\mathcal{X}$, f must attain the global minimum in $\mathcal{X}$.

In an attempt to alleviate the curse of dimensionality of generic global optimization, we focus on objective functions with '*low effective dimensionality*' [163], namely, those that

---

[12] Without loss of generality, we can assume that $\mathcal{X} = [-1, 1]^D$ as every rectangular domain can be scaled to the hypercube $\mathcal{X}$.

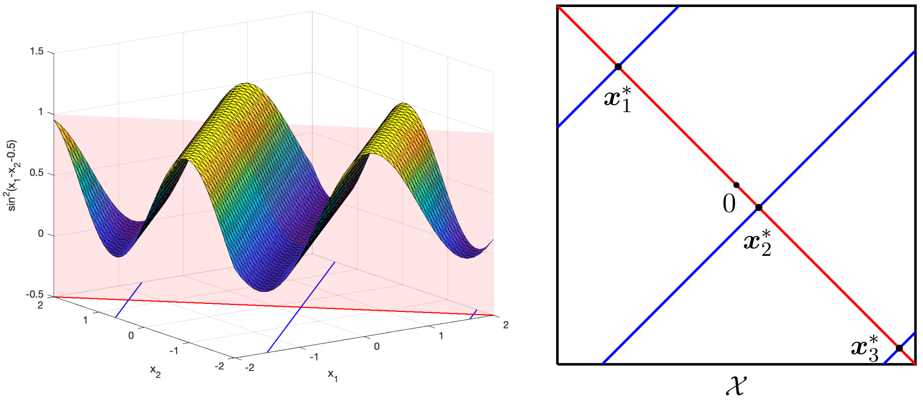


Figure 1.4: The figure on the left depicts the function in (1.12) and the figure on the right represents the domain space $\mathcal{X}$ of the function. The red line is the embedded line $(1\ -1)^T y$. The blue lines represent the set of global minimizers, which is comprised of the sets G_1^* (top left corner), G_2^* (middle), and G_3^* (bottom right corner).

only vary over a low-dimensional *effective subspace* (which may not necessarily be aligned with standard axes), and remain constant along its orthogonal complement. These functions are also known as objectives with '*active subspaces*' [37] or '*multi-ridge*' [55, 156]. They are frequently encountered in applications, typically when tuning (over)parametrized models and processes, such as in hyper-parameter optimization for neural networks [17], heuristic algorithms for combinatorial optimization problems [76], complex engineering and physical simulation problems [37] as in climate modelling [86], and policy search and dynamical system control [57, 174].

The presence of effective subspace, when known, allows us to reduce the dimension of the function and thereby reduce the computational costs of solving the optimization problem. We illustrate the core idea underlying the dimensionality reduction method on an illustrative example.

Example 1.1. Consider the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{x}} \ & \sin^2(x_1 - x_2 - 0.5), \\ \text{s.t.} \ & \boldsymbol{x} \in \mathcal{X} = [-2, 2]^2. \end{aligned} \tag{1.12}$$

The global minima of the problem can be found analytically. By solving $\sin^2(x_1 - x_2 - 0.5) = 0$, we find that the set of global minimizers comprises of the following three sets

$$G_1^* = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix} t - \begin{bmatrix} 0 \\ 0.5 - \pi \end{bmatrix} : -2 \leq t \leq 2.5 - \pi \right\}, \tag{1.13}$$

$$G_2^* = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix} t - \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} : -1.5 \leq t \leq 2 \right\}, \tag{1.14}$$

$$G_3^* = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix} t - \begin{bmatrix} 0 \\ 0.5 + \pi \end{bmatrix} : -1.5 + \pi \leq t \leq 2 \right\}. \tag{1.15}$$

Note that each G_i^* corresponds to a distinct line of global minimizers along which the function is constant and at which it attains its global minimum (see Figure 1.4).

We can reduce the dimension of the problem by one by parametrizing $x_1 = y$ and $x_2 = -y$; this gives us the following one-dimensional problem:

$$\begin{aligned} \min_{y} \quad & \sin^2(2y - 0.5), \\ \text{s.t.} \quad & (1 \ -1)^T y \in \mathcal{X}. \end{aligned}$$

The solutions are achieved at points $y^* = \pi i/2 + 0.25$, $i = 0, \pm 1$. Geometrically, the parametrization is equivalent to the embedding of the line $(1 \ -1)^T$ into the two-dimensional space and the search for the global minima along this line. The solutions in the reduced space correspond to the intersection points between $(1 \ -1)^T$ and the lines of global minimizers. The three solutions in the original space are points $\boldsymbol{x}_i^* = (\pi k + 0.25) \cdot (1 \ -1)^T$ for $i = 1, 2, 3$.

Note that not every parametrization would work. For example, if the embedding is parallel to the line of the minimisers, i.e., if $\boldsymbol{x} = (1 \ \ 1)^T y$, the embedding fails to recover the global minimum of the original problem.

1.2.1 Reduced problems with random embeddings

As Example 1.1 illustrates, when the objective has low effective dimensionality and the effective subspace of variation is known, it is possible to cast (UP) ((P)) into a lower-dimensional problem — which has the same global minimum f_U^* (f^*) — by restricting it to and solving (UP) ((P)) only within this important subspace. Typically, however, the effective subspace is unknown[13], and random embeddings have been proposed to reduce the size of (UP) ((P)) and hence the cost of its solution, while attempting to preserve the problem's original global minimum values. In this thesis, we investigate the following two reduced randomized problems

$$\begin{aligned} \min_{\boldsymbol{y}} \quad & f(\boldsymbol{A}\boldsymbol{y} + \boldsymbol{p}) \\ \text{s. t.} \quad & \boldsymbol{y} \in \mathcal{Y} = [-\delta, \delta]^d, \end{aligned} \qquad \text{(RP)} \qquad \begin{aligned} \min_{\boldsymbol{y}} \quad & f(\boldsymbol{A}\boldsymbol{y} + \boldsymbol{p}) \\ \text{s. t.} \quad & \boldsymbol{A}\boldsymbol{y} + \boldsymbol{p} \in \mathcal{X} = [-1, 1]^D, \end{aligned} \qquad \text{(RP}\mathcal{X}\text{)}$$

where $\boldsymbol{A}$ is a $D \times d$ Gaussian random matrix (see Definition 2.1) with $d \ll D$, and where $\boldsymbol{p}$ and $\delta > 0$ are user-defined. We use (RP) to tackle the unconstrained problem (UP) and use (RP$\mathcal{X}$) to tackle the constrained problem (P).

Remark 1.2. If $\boldsymbol{y}^*$ is a global solution to (RP) or (RP$\mathcal{X}$), one can recover a corresponding minimizer $\boldsymbol{x}^*$ in the original dimension by setting $\boldsymbol{x}^* = \boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}$.

---

[13] For example, in applications where f is black-box.

We say (RP) ((P)) is successful if its global minimum coincides with the global minimum f_U^* (f^*) of (UP) ((P)). We define this notion formally below.

Definition 1.3. We say that (RP) is *successful* if there exists $\boldsymbol{y}^* \in \mathbb{R}^d$ such that $f(\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}) = f_U^*$ and $\boldsymbol{y}^* \in \mathcal{Y}$.

Definition 1.4. We say that (RP$\mathcal{X}$) is *successful* if there exists $\boldsymbol{y}^* \in \mathbb{R}^d$ such that $f(\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}) = f^*$ and $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} \in \mathcal{X}$.

As noted in Example 1.1, not all embeddings can successfully recover the global minimum of the original problem. Therefore, we expect that the reduced problems (RP) and (RP$\mathcal{X}$) will be successful only for certain matrices $\boldsymbol{A}$. Furthermore, given that $\boldsymbol{A}$'s are initialized randomly, (RP) and (RP$\mathcal{X}$) will be successful in probability. In this thesis, we study the probabilities of success of the reduced problems (RP) and (RP$\mathcal{X}$). In particular, we investigate how these probabilities are affected by the user-chosen parameters of the reduced problems and characteristics intrinsic to the objective function (e.g., orientation of the effective subspace).

Description of the reduced random problem (RP)

As mentioned above, with formulation (RP) we aim to recover the global minimum of (UP).

Remark 1.5. Note that, unlike (UP), (RP) has (box) constraints, which are typically imposed to make the approach practical (i.e. to avoid unrealistic searches over infinite domains).

Problem (RP) involves three user-chosen parameters $\boldsymbol{p}$, d and δ. These parameters determine the probability of success of (RP) in recovering f_U^*. It is clear, for example, that for larger values of d and/or δ the probability of success of (RP) is higher. On the other hand, larger values of d — the dimension of (RP) — and/or δ — the half-length of the domain — demand more computational resources to solve (RP). Therefore, a careful calibration of these two parameters is needed to ensure that (RP) is successful for most embeddings, at the same time being capable to converge to the solution within the computational budget. Relying on the derivation of the distribution of a reduced minimizer presented in Chapter 2, we analyse and discuss the effects of the three parameters on the success of the reduced problem in Chapter 3. We will see later that, with the right choices of the parameters, a single solve of (RP) is often sufficient to find the value of f_U^*.

Description of the reduced random problem (RP$\mathcal{X}$)

To tackle the constrained problem (P), we turn to formulation (RP$\mathcal{X}$).

Just like for (RP), the right choice of parameters $\boldsymbol{p}$ and d is crucial to the success of (RP$\mathcal{X}$). To analyse the probability of success of (RP$\mathcal{X}$) and its relationship to $\boldsymbol{p}$, d and D, we will rely on the analysis of the distribution of a reduced minimizer presented in Chapter 2.

While problem (RP) is often successful with only one embedding, (RP$\mathcal{X}$) may require multiple embeddings[14] to converge to the global minimum. Multiple instances of (RP$\mathcal{X}$) can be solved simultaneously using parallel computing. This approach can speed up the overall procedure but lacks interdependence between the different solves. One can also perform multiple embeddings sequentially using information from the past solves to perform the next iteration. We explore the formulation (RP$\mathcal{X}$) in Chapter 4.

1.2.2 Existing relevant literature

The scalability challenges of Bayesian Optimization (BO) algorithms for generic black-box functions have prompted research into improving efficiency of this class of methods for functions with special structure. Different structural assumptions on the objective have been analysed for BO, such as additivity or (partial) separability, which assumes that the objective function can be represented as the sum of smaller-dimensional functions with non-overlapping variables [164, 81, 97] or with overlapping ones [137].

Another popular structural assumption is the low-effective dimensionality of the objective, which is the focus of the thesis. In its simplest form, this considers the effective subspace to be aligned with the coordinate axes, which is equivalent to the presence of redundant variables [33, 15]. More generally, the optimization of functions that are constant along *arbitrary* linear subspaces has been addressed using BO methods in [43, 163, 60, 48], and extended to other problem and algorithm classes such as derivative-free optimization [132], multi-objective optimization [131] and evolutionary methods [141]. As the effective subspace is generally unknown, some existing approaches learn the effective subspace beforehand [43, 48, 55, 156], while others estimate it during the optimization, updating the estimate as new information becomes available on the objective function [34, 41, 60, 174]. We focus here on an alternative approach, bypassing the subspace learning phase, and optimizing directly over random low-dimensional subspaces, as proposed in [19, 20, 85, 163].

Our work was particularly inspired by Wang et al. [163]'s REMBO algorithm — a BO framework that uses the reduced problem (RP) with $\boldsymbol{p} = \boldsymbol{0}$ to solve[15] the box-constrained

---

[14] In other words, we solve (RP$\mathcal{X}$) numerous times each time with a different Gaussian matrix $\boldsymbol{A}$.

[15] Note that, in this thesis, we are using (RP) to solve the unconstrained problem (UP).

problem (P). They find that the size of $\mathcal{Y}$ is the primary factor in determining the success (or failure) of the reduced problem, and quantify the probability of success of (RP) for the case when the effective subspace is aligned with the coordinate axes and when $d = d_e$, where d_e denotes the dimension of the effective subspace (see [163, Theorem 3]). The theoretical analysis in [163] has been extended by Sanyang and Kabán [141], where the probability of success of (RP) is quantified for a wider range of values of d; an algorithm, called REMEDA, is also proposed in [141] that uses Gaussian random embeddings in the framework of evolutionary methods for high-dimensional unconstrained global optimization.

A challenge of (RP) when applied to the box-constrained problem (P) is that — even when (RP) is successful — the high-dimensional image $\boldsymbol{A}\boldsymbol{y} \in \mathbb{R}^D$ of a point $\boldsymbol{y} \in \mathcal{Y}$ may be outside the feasible set $\mathcal{X}$. To remedy this, Wang et al. [163] endow REMBO with an additional step that projects $\boldsymbol{A}\boldsymbol{y}^*$ onto $\mathcal{X}$. However, they observe that using a classical kernel (such as the squared exponential kernel, see (1.6)) directly on the low-dimensional domain $\mathcal{Y}$ may lead to an over-exploration of the regions on which the projection map onto $\mathcal{X}$ is not injective. The design of kernels avoiding this over-exploration has been tackled in [19, 20]. Binois et al. [19] propose a new kernel k_Ψ which has the benefit of being low-dimensional while avoiding the over-exploratory tendency of $k_\mathcal{Y}$. In [20], Binois et al. also propose a new mapping γ (instead of $p_\mathcal{X}$) and define $\mathcal{Y}$ and new kernels based on this new mapping. Nayebi et al. [114] circumvent the projection step by replacing the Gaussian random embeddings of (RP) by random embeddings defined using hashing matrices, and choose $\mathcal{Y} = [-1, 1]^d$. This choice guarantees that any solution of the low-dimensional problem provides an admissible solution for the full-dimensional problem in the case $\mathcal{X} = [-1, 1]^D$.

The infeasibility challenge of (RP) can be avoided altogether if (RP) is replaced by the formulation (RP$\mathcal{X}$). Using (RP$\mathcal{X}$) for solving (P), however, poses a new challenge of a low probability of success. This was pointed out by Letham et al. [93], where they consider a formulation similar to (RP$\mathcal{X}$) and estimate the probability of success using Monte-Carlo simulations. To compensate for the reduced success probability of a single application of (RP$\mathcal{X}$), one can solve (RP$\mathcal{X}$) multiple times each time with a different embedding, thus, covering more space in $\mathcal{X}$ and increasing the overall chance of success. Applications of iterative subspace embeddings have been suggested, for example, in Qian et al. [132] for unconstrained derivative-free optimization and, in Kirschner et al. [85] for constrained BO optimization, where one-dimensional subspace embeddings are considered. Qian et al. [132] also introduce the notion of approximate low effective subspace, whereas most of the previous work concerns objective functions with exact low effective subspaces.

1.3 Thesis contributions and outline

In this thesis, we investigate two general random embeddings frameworks for unconstrained and box-constrained global optimization of functions with low effective dimensionality, where we allow the effective subspace of the objective function to be arbitrary (not necessarily aligned with coordinate axes and not limited in its dimension d_e by problem constants). We significantly extend and improve theoretical analyses in existing literature, providing an in-depth investigation of the reduced problems (RP) and (RP$\mathcal{X}$) when Gaussian random embeddings are used. We also investigate their algorithmic potential in solving the original problems (UP) and (P) respectively, both theoretically and numerically.

In Chapter 2, for both (RP) and (RP$\mathcal{X}$), we define a (random) reduced minimizer $\boldsymbol{y}_2^*$ expressed in terms of the embedding matrix $\boldsymbol{A}$, a basis of the effective subspace and an arbitrary global minimizer $\boldsymbol{x}^*$ of (UP) or (P), depending on the problem being considered. In the literature (e.g. in [163, 141]), this reduced minimizer has played a key role in estimating the probability of success of (RP). While [163, 141] have only attempted to estimate the Euclidean norm of $\boldsymbol{y}_2^*$, we derive its exact distribution; this leads to more precise bounds for (RP) and a novel type of analysis to estimate the probability of success of (RP$\mathcal{X}$). Using properties of Gaussian and related distributions, we show that $\boldsymbol{y}_2^*$, when appropriately scaled, follows the inverse chi-squared distribution with $d - d_e + 1$ degrees of freedom (Theorem 2.27). Moreover, we derive the probability density function (p.d.f.) of $\boldsymbol{y}_2^*$ (Theorem 2.30) by first proving that it follows a spherical distribution.

Based on the distribution of the norm of this reduced minimizer, we estimate the probability of success of (RP) in Theorem 3.3 and Corollary 3.4. The latter result extends both [163, Theorem 3] and [141, Theorem 2] to arbitrary effective subspaces and any $d \geq d_e$, and establishes a notable and more precise trade-off between the success of (RP), δ (the size of the reduced domain $\mathcal{Y}$) and the embedding dimension d; thus allowing us to choose appropriate values for these parameters in the algorithm. Moreover, Corollary 3.4 implies that, under certain general assumptions, solving (RP) has no dependence on the ambient dimension D.

Using the distribution of the reduced minimizer $\boldsymbol{y}_2^*$, we also derive a lower bound on the probability of success of (RP$\mathcal{X}$) when $d \geq d_e$. To achieve this, we provide a sufficient condition for the success of (RP$\mathcal{X}$) that depends on a random vector $\boldsymbol{w}$, which in turn, is a function of the embedding matrix $\boldsymbol{A}$ and the reduced minimizer $\boldsymbol{y}_2^*$. We show that $\boldsymbol{w}$ follows a $(D - d_e)-$dimensional t-distribution with $d - d_e + 1$ degrees of freedom (Theorem 4.5), and provide a lower bound on the probability of success of (RP$\mathcal{X}$) in terms of the integral of the p.d.f. of $\boldsymbol{w}$ over a given closed domain (Corollary 4.6). In the case when the effective subspace is aligned with the coordinate axes, this closed domain simplifies to a $(D-d_e)-$dimensional box, and we provide an asymptotic expansion of the integral of the p.d.f. over the box (Theorems 4.13 and 4.14), when $D \to \infty$ (and d and d_e are fixed). Our theoretical analysis, backed by numerical testing, indicates that the probability of success of (RP$\mathcal{X}$) decreases with the dimension D of the original problem (P). However, in the case when the effective subspace is aligned with the coordinate axes, we show that it decreases at most algebraically with the ambient dimension D for some useful choices of $\boldsymbol{p}$.

In Chapter 5, we propose to extend the scope of (RP$\mathcal{X}$) to general objectives (that may not possess low effective dimensionality). We develop a novel analysis to bound the probability of success[16] of (RP$\mathcal{X}$) for Lipschitz continuous functions based on connections between (RP$\mathcal{X}$) and the field of conic integral geometry (see, e.g., [6, 124]). Our analysis, unlike (RP) or (RP$\mathcal{X}$) for functions with low effective dimensionality, is applicable for any embedding dimension $d \geq 1$. To understand relative performance of (RP$\mathcal{X}$), we compare (RP$\mathcal{X}$) to uniform sampling technique (uniform random search) by contrasting the bounds implied by the two approaches in the asymptotic regime.

We propose two algorithmic frameworks, namely, REGO (Random Embeddings for Global Optimization) that solves a single randomly-embedded reduced problem (RP) instead of (UP), and X-REGO for the constrained global optimization problem (P) that sequentially or in parallel solves multiple subproblems (RP$\mathcal{X}$), varying $\boldsymbol{A}$ and also possibly $\boldsymbol{p}$. These frameworks are compatible with any generic global optimization solver.

We use and validate our theoretical results related to (RP) by providing extensive numerical testing of REGO with three types of solvers: DIRECT, BARON, and (multi-start) KNITRO (see above or Section 3.3). We use 19 standard global optimization test problems to generate functions with effective dimensionality structure and of growing ambient dimension D. When comparing REGO with the direct optimization of the ensuing problems without embeddings, we find that REGO's performance is essentially independent of D for all three solvers and that it is successful in recovering the original global minimum in most cases with only one embedding[17]. We also test the robustness of REGO's performance to variations in algorithm parameters such as δ and d.

We dedicate Chapter 6 to the X-REGO algorithm, proving its global convergence to a set of approximate global minimizers of (P) with probability one, with linear rate in terms of the number of subproblems solved. This result requires mild assumptions on problem (P) (f is Lipschitz continuous and (P) admits a strictly feasible solution) and on the algorithm

---

[16] Here, success of (RP$\mathcal{X}$) means that the embedded subspace intersects the set of approximate global minimizers.

[17] Numerical results of REGO (as well as of X-REGO) assume that (an upper bound on) the true effective dimension d_e is known/available.

used to solve the reduced problem (namely, it must solve (RP$\mathcal{X}$) globally and approximately, to required accuracy at least with a certain probability), and allows a diverse set of possible choices of $\boldsymbol{p}$ (random, fixed, adaptive, deterministic). Our framework here can be viewed as a hybrid technique that merges multi-dimensional random search (random subspaces) with the power of global/local algorithms in low dimensions.

Also in Chapter 6, we provide an extensive numerical comparison of several variants of X-REGO on the same test set of functions with low effective dimensionality used for REGO, employing the above mentioned three global solvers as well as local KNITRO. We find that X-REGO variants show significantly improved scalability with most solvers, as the ambient problem dimension grows, compared to directly using the respective solvers on the test set. Notable efficiency was obtained in particular when local KNITRO was used to solve the subproblems and the points $\boldsymbol{p}$ were updated to the 'best' point (with the smallest value of f) found so far.

Outline. We begin Chapter 2 by recalling technical definitions and tools related to Gaussian matrix and related distributions. We then formally define the notion of functions with low effective dimensionality and describe the geometry of random embeddings in the context of the two reduced problems. For both reduced problems, we also define a random reduced minimizer and derive its distribution.

In Chapter 3, based on the distribution of the reduced minimizer, we derive a lower bound on the probability of success of (RP) and compare this bound with existing bounds from the literature. Here, we also introduce the REGO framework and conduct numerical experiments. The content of this chapter and Chapter 2 has been published in [29].

We dedicate Chapter 4 to derivations of lower bounds for the success of (RP$\mathcal{X}$). We start the chapter with a derivation of the distribution of the image of the reduced minimizer in the ambient space. Using this result, we then derive a general lower bound for the success probability of (RP$\mathcal{X}$). This bound leads to several (interpretable/quantifiable) bounds. For example, we show that the probability of success of (RP$\mathcal{X}$) is positive and derive bounds for the case when the effective subspace is aligned with coordinate axes. Here, we also define the notion of approximate success of (RP$\mathcal{X}$) and show that the probability of (RP$\mathcal{X}$) being approximately successful is bounded away from zero uniformly in $\boldsymbol{p}$.

In Chapter 5, we consider (RP$\mathcal{X}$) but now without the assumption that the objective has low effective dimensionality. In our analysis, assuming that the objective is Lipschitz continuous, we bound the probability of (RP$\mathcal{X}$) capturing the set of approximate global minimizers relying on the earlier description of key results from conic integral geometry. At the end of the chapter, we compare (RP$\mathcal{X}$) to a simple random search.

We end our thesis with Chapter 6, where we introduce and analyse the X-REGO framework. We start the chapter with the description of the framework. This will be followed by the global convergence proof and its implications for objective functions with and without low effective dimensionality. At the end of the chapter, we present our numerical experiments with X-REGO. The content of this chapter and Chapter 4 has been submitted for publication as [32].

Chapter 3–6 end with summaries that review the main contributions of the chapters as well as point to any limitations and potential future directions of the respective work.

1.4 Notation

We use bold capital letters for matrices ($\boldsymbol{A}$) and bold lowercase letters ($\boldsymbol{a}$) for vectors. In particular, $\boldsymbol{I}_D$ is the $D \times D$ identity matrix and $\mathbf{0}_D$, $\mathbf{1}_D$ (or simply $\mathbf{0}$, $\mathbf{1}$) are the D-dimensional vectors of zeros and ones, respectively. We write a_i to denote the ith entry of $\boldsymbol{a}$ and write $\boldsymbol{a}_{i:j}$, $i < j$, for the vector $(a_i \; a_{i+1} \cdots a_j)^T$. We let range($\boldsymbol{A}$) denote the linear subspace spanned in $\mathbb{R}^D$ by the columns of $\boldsymbol{A} \in \mathbb{R}^{D \times d}$. We write $\| \cdot \|$ and $\| \cdot \|_\infty$ for the Euclidean norm and the infinity norm, respectively. Where emphasis is needed, for the Euclidean norm we also use $\| \cdot \|_2$.

Given two random variables (vectors) x and y ($\boldsymbol{x}$ and $\boldsymbol{y}$), the expression $x \overset{law}{=} y$ ($\boldsymbol{x} \overset{law}{=} \boldsymbol{y}$) means that x and y ($\boldsymbol{x}$ and $\boldsymbol{y}$) have the same distribution. We reserve the letter $\boldsymbol{A}$ for a $D \times d$ Gaussian random matrix (see Definition 2.1) and write χ^2_n to denote a chi-squared random variable with n degrees of freedom (see Definition 2.5).

Given a point $\boldsymbol{a} \in \mathbb{R}^D$ and a set S of points in $\mathbb{R}^D$, we write $\boldsymbol{a} + S$ to denote the set $\{\boldsymbol{a} + \boldsymbol{s} : \boldsymbol{s} \in S\}$. Given functions $f(x) : \mathbb{R} \to \mathbb{R}$ and $g(x) : \mathbb{R} \to \mathbb{R}^+$, we write $f(x) = \Theta(g(x))$ as $x \to \infty$ to denote the fact that there exist positive reals M_1, M_2 and a real number x_0 such that, for all $x \geq x_0$, $M_1 g(x) \leq |f(x)| \leq M_2 g(x)$.

Chapter 2

Functions with low effective dimensionality: tools, formulations and characterizations

The goal of this chapter is to lay out the technical foundation for the upcoming chapters. We begin this chapter with technical definitions and results that pertain to Gaussian and related distributions. Then, in Section 2.2, we formally define the notion of functions with low effective dimensionality. This will be followed by a geometric description of random embeddings technique in Section 2.3 and derivations of various probabilistic results related to the reduced problems (RP) and (RP$\mathcal{X}$) in Section 2.4.

2.1 Gaussian matrix distribution: related definitions and tools

We start the chapter with the technical definitions and results related to Gaussian matrix distribution.

2.1.1 Gaussian random matrices

We begin with the definition of a Gaussian random matrix.

Definition 2.1. (Gaussian matrix, see [70, Definition 2.2.1]) A Gaussian (random) matrix is a matrix $\boldsymbol{M} = (m_{ij})$, where the entries $m_{ij} \sim \mathcal{N}(0,1)$ are independent (identically distributed) standard normal variables.

Gaussian matrices have been well-studied with many results available at hand. Here, we mention a few key properties of Gaussian matrices which we will use in the analysis. For a collection of results pertaining to Gaussian matrices and other related distributions the reader is directed to [70, 159].

Gaussian matrices are famous for their orthogonal invariance property.

Theorem 2.2. *[70, Theorem 2.3.10] Let* $\boldsymbol{M}$ *be an* $M_1 \times M_2$ *Gaussian random matrix. If* $\boldsymbol{U} \in \mathbb{R}^{M_1 \times r}$, $M_1 \geq r$, *and* $\boldsymbol{V} \in \mathbb{R}^{M_2 \times q}$, $M_2 \geq q$, *are orthogonal, then* $\boldsymbol{U}^T\boldsymbol{M}\boldsymbol{V}$ *is a Gaussian random matrix.*

In the analysis, we will also need the following result.

Theorem 2.3. *[70, Theorem 2.3.15] Let* $\boldsymbol{M}$ *be an* $M_1 \times M_2$ *Gaussian random matrix, and let* $\boldsymbol{X} \in \mathbb{R}^{r \times M_1}$ *and* $\boldsymbol{Y} \in \mathbb{R}^{q \times M_1}$ *be given matrices. Then,* $\boldsymbol{X}\boldsymbol{M}$ *and* $\boldsymbol{Y}\boldsymbol{M}$ *are independent if and only if* $\boldsymbol{X}\boldsymbol{Y}^T = \boldsymbol{0}$.

A related notion that plays an important role in the study of Gaussian matrices is the Wishart distribution represented by matrix $\boldsymbol{M}^T\boldsymbol{M}$ (or $\boldsymbol{M}\boldsymbol{M}^T$), where $\boldsymbol{M}$ is an overdetermined (undetermined) Gaussian random matrix. A Wishart matrix is positive definite, and hence nonsingular, with probability 1.

Theorem 2.4. *(see [70, Theorem 3.2.1]) Let* $\boldsymbol{M}$ *be an* $M_1 \times M_2$ *Gaussian random matrix,* $M_1 \geq M_2$. *Then, the Wishart matrix* $\boldsymbol{M}^T\boldsymbol{M}$ *is positive definite with probability 1.*

2.1.2 Chi-squared distribution

A closely related distribution to Gaussian is the chi-squared. A chi-squared random variable will feature prominently in our theoretical results. Let us first recall the definition of a chi-squared random variable.

Definition 2.5 (Chi-squared random variable). Given a collection $Z_1, Z_2, \ldots, Z_N$ of N independent standard normal variables, a random variable $X = Z_1^2 + Z_2^2 + \cdots + Z_N^2$ is said to follow the chi-squared distribution with N degrees of freedom. We denote this by $X \sim \chi_N^2$. The probability density function (p.d.f.) $h(N,x)$ and the cumulative density function (c.d.f.) $H(N,x)$ of X are given by (see, e.g., [1, formula 26.4.1] and [122]):

$$h(N,x) = \frac{1}{2^{N/2}\Gamma(N/2)} x^{N/2-1} e^{-x/2} \text{ and } H(N,x) = \frac{\gamma(N/2, x/2)}{\Gamma(N/2)}, \tag{2.1}$$

for $x > 0$, where Γ is the gamma function, which is defined as

$$\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du \quad \text{for } a > 0, \tag{2.2}$$

and where $\gamma(a,t)$ is the lower incomplete gamma function (see [120]) defined as

$$\gamma(a,t) = \int_0^t u^{a-1} e^{-u} du \quad \text{for } a > 0,\, t > 0. \tag{2.3}$$

The following lemmas provide notable relationships between a Gaussian matrix and chi-squared random variable.

Lemma 2.6. *([70, Corollary 3.3.13.1]) Let $\boldsymbol{M}$ be an $M_1 \times M_2$ Gaussian matrix, $M_1 \geq M_2$, and let $\boldsymbol{z} \in \mathbb{R}^{M_2}$ be a fixed non-zero vector. Then,*

$$\frac{\|\boldsymbol{z}\|^2}{\boldsymbol{z}^T(\boldsymbol{M}^T\boldsymbol{M})^{-1}\boldsymbol{z}} \sim \chi^2_{M_1-M_2+1}.$$

Theorem 2.7. *([70, Theorem 3.3.12]) Let $\boldsymbol{M}$ be an $M_1 \times M_2$ Gaussian matrix, $M_1 \geq M_2$, and $\boldsymbol{y}$ be an $M_2 \times 1$ random vector distributed independently of $\boldsymbol{M}^T\boldsymbol{M}$, and $\mathbb{P}[\boldsymbol{y} \neq \boldsymbol{0}] = 1$. Then,*

$$\frac{\boldsymbol{y}^T\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y}} \sim \chi^2_{M_1}$$

and is independent of $\boldsymbol{y}$.

2.1.3 Miscellaneous probability distributions

Here, we provide definitions on three distributions, namely, inverse chi-squared, multivariate t- and F-distributions. We will encounter the inverse chi-squared random variable later in this chapter, in Chapters 3 and 4, whilst the multivariate t and F distributions will come up in Chapter 4.

Definition 2.8 (The inverse chi-squared random variable). Given $X \sim \chi^2_N$, a random variable $Y = 1/X$ is said to follow the inverse chi-squared distribution with N degrees of freedom. We denote this by $Y \sim 1/\chi^2_N$. The p.d.f. of Y is given by (see [92, A5]):

$$h(N, y) = \frac{1}{2^{N/2}\Gamma(N/2)} y^{-N/2-1} e^{-1/(2y)} \tag{2.4}$$

for $y > 0$.

A short lemma that provides the expectation of the inverse chi-squared random variable is given next.

Lemma 2.9. *(see [92, A5]) Let $Y \sim 1/\chi^2_N$. Then,*

$$\mathbb{E}[Y] = \frac{1}{N-2}$$

provided that $N > 2$.

Let us now define multivariate t-distribution.

Definition 2.10 (Multivariate $\boldsymbol{t}$-distribution). A q-dimensional random variable $\boldsymbol{t}$ is said to have t-distribution with parameters ν and $\boldsymbol{\Sigma}$ if its joint p.d.f. is given by (see [70, Chapter 4])

$$h(\boldsymbol{t}) = \frac{1}{(\pi\nu)^{q/2}} \left[\frac{\Gamma\left(\frac{q+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right] \det(\boldsymbol{\Sigma})^{-1/2} \left(1 + \frac{1}{\nu}\boldsymbol{t}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{t}\right)^{-(q+\nu)/2}, \ \boldsymbol{t} \in \mathbb{R}^q. \tag{2.5}$$

The t-distributed random vector can be represented as (an appropriately scaled) product of a Gaussian vector and an inverse chi-squared random variable. For example, note the following result.

Theorem 2.11. *(see [70, eq. (4.1.1)]) Let* $Z_1, Z_2, \dots, Z_N$ *be* N *independent standard normal variables, let* X *follow the chi-squared distribution with* ν *degrees of freedom and let* $\boldsymbol{t}$ *follow an* N*-dimensional t-distribution with parameters* ν *and* $\sigma^2 \boldsymbol{I}$*. Then, we have*

$$\boldsymbol{t} \overset{law}{=} \sqrt{\frac{\nu\sigma^2}{X}} \begin{pmatrix} Z_1 \\ \vdots \\ Z_N \end{pmatrix}. \tag{2.6}$$

Lastly, we define F-distribution providing its probability density function.

Definition 2.12 (F-distribution). Let $W_1 \sim \chi^2_{N_1}$ and $W_2 \sim \chi^2_{N_2}$ be independent. A random variable X is said to follow an F-distribution with degrees of freedom N_1 and N_2 if

$$X \sim \frac{W_1/N_1}{W_2/N_2}.$$

We denote this by $X \sim F(N_1, N_2)$. The p.d.f. of X is given by (see [92, A.19])

$$h(x) = \frac{\Gamma(\frac{N_1+N_2}{2})}{\Gamma(\frac{N_1}{2})\Gamma(\frac{N_2}{2})} \left(\frac{N_1}{N_2}\right)^{N_1/2} x^{N_1/2-1} \left(1 + \frac{N_1}{N_2}x\right)^{-\frac{N_1+N_2}{2}} \quad \text{for } x > 0. \tag{2.7}$$

2.1.4 Spherical distribution

Spherical distributions are a class of probability distributions that generalize Gaussian distribution. The generalization is based on the orthogonal invariance property of Gaussian distribution.

Definition 2.13. A $q \times 1$ random vector $\boldsymbol{x}$ is said to have a spherical distribution if for every orthogonal $q \times q$ matrix $\boldsymbol{Q}$,

$$\boldsymbol{Q}\boldsymbol{x} \overset{law}{=} \boldsymbol{x}.$$

Interestingly, it is sufficient to know the distribution of the Euclidean norm of a spherically distributed random vector to be able to derive its p.d.f.

Theorem 2.14. *([51, Corollary, p. 30; Theorem 2.3] If* $q \times 1$ *random vector* $\boldsymbol{x}$ *has a spherical distribution, then*

$$\boldsymbol{x} \overset{law}{=} r\boldsymbol{u},$$

where $\boldsymbol{u}$ *is distributed uniformly on the* $(q-1)$*-dimensional unit sphere and* r *is a univariate random variable independent of* $\boldsymbol{u}$*. If, in addition,* $\mathbb{P}[\boldsymbol{x} = \boldsymbol{0}] = 0$*, then*

$$\|\boldsymbol{x}\| \overset{law}{=} r \text{ and } \|\boldsymbol{x}\|^{-1}\boldsymbol{x} \overset{law}{=} \boldsymbol{u}.$$

Theorem 2.15. *(see [71, Theorem 2.1.]) Let* $\boldsymbol{x} \overset{law}{=} r\boldsymbol{u}$ *be a spherically distributed* $q \times 1$ *random vector with* $\mathbb{P}[\boldsymbol{x} = \boldsymbol{0}] = 0$*, where* $r = \|\boldsymbol{x}\|$ *is independent of* $\boldsymbol{u} = \boldsymbol{x}/\|\boldsymbol{x}\|$ *and has p.d.f.* $h(\hat{r})$*. Then, p.d.f.* $g(\hat{\boldsymbol{x}})$ *of* $\boldsymbol{x}$ *is given by*

$$g(\hat{\boldsymbol{x}}) = \frac{\Gamma(q/2)}{2\pi^{q/2}} h(\|\hat{\boldsymbol{x}}\|)\|\hat{\boldsymbol{x}}\|^{1-q}.$$

For more details regarding spherical distributions refer to [51, 70, 18].

2.2 Functions with low effective dimensionality

Functions with low effective dimensionality can be defined in at least two ways [55, 163]. We will work with the definition given in terms of linear subspaces, provided in [163].

Definition 2.16 (Functions with low effective dimensionality). A function $f : \mathbb{R}^D \to \mathbb{R}$ has effective dimension d_e if there exists a linear subspace $\mathcal{T}$ of dimension d_e such that for all vectors $\boldsymbol{x}_\top$ in $\mathcal{T}$ and $\boldsymbol{x}_\perp$ in $\mathcal{T}^\perp$ (the orthogonal complement of $\mathcal{T}$), we have

$$f(\boldsymbol{x}_\top + \boldsymbol{x}_\perp) = f(\boldsymbol{x}_\top), \tag{2.8}$$

and d_e is the smallest integer satisfying (2.8).

The linear subspaces $\mathcal{T}$ and $\mathcal{T}^\perp$ are called the *effective* and *constant* subspaces of f, respectively. The following assumption on the function f will be frequently used in the thesis.

Assumption LowED. The function $f : \mathbb{R}^D \to \mathbb{R}$ is continuous and has effective dimensionality d_e such that $d_e < D$, with effective subspace[1] $\mathcal{T}$ and constant subspace $\mathcal{T}^\perp$ spanned by the columns of the orthonormal matrices $\boldsymbol{U} \in \mathbb{R}^{D\times d_e}$ and $\boldsymbol{V} \in \mathbb{R}^{D\times(D-d_e)}$, respectively. We write $\boldsymbol{x}_\top = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}$ and $\boldsymbol{x}_\perp = \boldsymbol{V}\boldsymbol{V}^T\boldsymbol{x}$, the unique Euclidean projections of any vector $\boldsymbol{x} \in \mathbb{R}^D$ onto $\mathcal{T}$ and $\mathcal{T}^\perp$, respectively.

2.2.1 Definitions relating to unconstrained problem (UP)

Recalling the definition of the unconstrained problem (UP) on page 18, we define the set of global minimizers

$$\mathcal{G}_U := \{\boldsymbol{x} \in \mathbb{R}^D : f(\boldsymbol{x}) = f_U^*\}. \tag{2.9}$$

Note that, for any $\boldsymbol{x}^* \in \mathcal{G}_U$ with Euclidean projection $\boldsymbol{x}_\top^*$ on the effective subspace $\mathcal{T}$, and for any $\tilde{\boldsymbol{x}} \in \mathcal{T}^\perp$, we have

$$f(\boldsymbol{x}^*) = f(\boldsymbol{x}_\top^* + \tilde{\boldsymbol{x}}) = f(\boldsymbol{x}_\top^*). \tag{2.10}$$

---

[1] Note that $\mathcal{T}$ in Assumption LowED may not be aligned with the standard/coordinate axes.

It is important to note that $\boldsymbol{x}_{\top}^{*}$ may not be unique, in fact, there can be infinitely many points $\boldsymbol{x}_{\top}^{*}$ in $\mathcal{X}$ satisfying the above definition. To understand why this might be the case refer to Example 1.1. The objective function $f = \sin^2(x_1 - x_2 - 0.5)$ has effective dimensionality 1 with $\mathcal{T}$ spanned by $(1 \;\; -1)^T$. The function has infinitely many $\boldsymbol{x}_{\top}^{*}$ in $\mathbb{R}^D$; this is due to the periodic nature of the sinus function — and not because f has low effective dimensionality. In this example, there are three points inside $\mathcal{X}$ that satisfy the definition of $\boldsymbol{x}_{\top}^{*}$; they are given by $\boldsymbol{x}_1^*$, $\boldsymbol{x}_2^*$ and $\boldsymbol{x}_3^*$ (see Figure 1.4). Compare this to $f = (x_1 - x_2 - 0.5)^2$. This function also has low effective dimensionality, but a unique $\boldsymbol{x}_{\top}^{*}$ given by $(0.25 \;\; -0.25)^T$.

Definition 2.17. Suppose Assumption LowED holds. For any $\boldsymbol{x}^* \in \mathcal{G}_U$, we define $\mathcal{G}_U^* := \boldsymbol{x}_{\top}^{*} + \mathcal{T}^{\perp}$ — the $(D - d_e)$-dimensional affine subspace of $\mathcal{G}_U$ that contains $\boldsymbol{x}_{\top}^{*} = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}^*$.

Remark 2.18. Note that if there is a unique $\boldsymbol{x}_{\top}^{*}$ then $\mathcal{G}_U = \mathcal{G}_U^*$. If there are multiple $\boldsymbol{x}_{\top}^{*}$'s then the set $\mathcal{G}_U$ is a union of the $(D - d_e)$-dimensional affine subspaces $\mathcal{G}_U^*$'s, each containing one particular $\boldsymbol{x}_{\top}^{*}$. For visualization of $\mathcal{G}_U^*$, refer to Figure 1.4 in Example 1.1, and imagine that there are no box constraints, so that the blue line segments depicted in the figure extend to infinity, i.e., become lines. These three lines are three different $\mathcal{G}_U^*$'s.

2.2.2 Definitions relating to constrained problem (P)

Likewise, for constrained problem (P) on page 18, we define[2]

$$\mathcal{G} := \{\boldsymbol{x} \in \mathbb{R}^D : f(\boldsymbol{x}) = f^*\} \tag{2.11}$$

and additionally, the set of feasible minimizers,

$$G := \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) = f^*\}. \tag{2.12}$$

Note that, for any $\boldsymbol{x}^* \in \mathcal{G}$ with Euclidean projection $\boldsymbol{x}_{\top}^{*}$ on the effective subspace $\mathcal{T}$, and for any $\tilde{\boldsymbol{x}} \in \mathcal{T}^{\perp}$, (2.10) holds.

Remark 2.19. For $\boldsymbol{x}^* \in G$, $\boldsymbol{x}_{\top}^{*}$ may lie outside $\mathcal{X}$.

Definition 2.20. Suppose Assumption LowED holds. For any $\boldsymbol{x}^* \in \mathcal{G}$, we define $\mathcal{G}^* := \boldsymbol{x}_{\top}^{*} + \mathcal{T}^{\perp}$ — the affine subspace of $\mathcal{G}$ that contains $\boldsymbol{x}_{\top}^{*} = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}^*$ and $G^* := \{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{x} \in \boldsymbol{x}_{\top}^{*} + \mathcal{T}^{\perp}\}$ — the simply connected subset of G that is contained in $\mathcal{G}^*$.

Same facts stated in Remark 2.18 hold for $\mathcal{G}$ and $\mathcal{G}^*$. We also add here that set G is a union of simply connected sets, each corresponding to a particular $\boldsymbol{x}_{\top}^{*}$, and if $\boldsymbol{x}_{\top}^{*}$ is unique then $G = \{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{x} \in \boldsymbol{x}_{\top}^{*} + \mathcal{T}^{\perp}\}$. For illustration, refer to Example 1.1, where we have

---

[2] Note that in general $\mathcal{G}_U \neq \mathcal{G}$. These sets are equal if all affine subspaces of global minimizers of (UP) pass through $\mathcal{X}$.

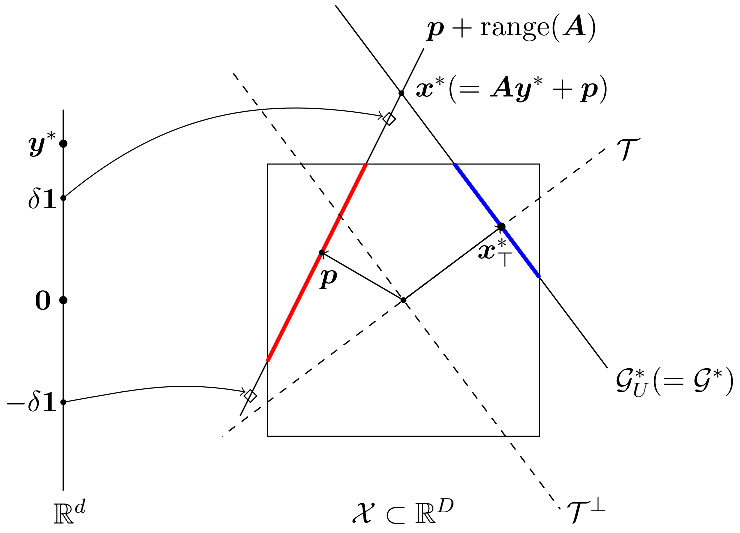

$\boldsymbol{p} + \text{range}(\boldsymbol{A})$
$\boldsymbol{x}^*(= \boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p})$
$\boldsymbol{y}^*$
$\mathcal{T}$
$\delta\boldsymbol{1}$
$\boldsymbol{x}^*_\top$
$\boldsymbol{p}$
$\boldsymbol{0}$
$\mathcal{G}^*_U(= \mathcal{G}^*)$
$-\delta\boldsymbol{1}$
$\mathbb{R}^d$
$\mathcal{X} \subset \mathbb{R}^D$
$\mathcal{T}^\perp$

Figure 2.1: Abstract illustration of the embedding of an affine d-dimensional subspace $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ into $\mathbb{R}^D$. For the sake of illustration, we assume that $\mathcal{G}^*_U = \mathcal{G}^*$. The red line represents the set of solutions along $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ that are contained in $\mathcal{X}$ and the blue line represents the set G^* (see Definition 2.20). (RP$\mathcal{X}$) is successful when the red and blue lines intersect. (RP) is successful if the line segment between two diamonds along $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ contains any point of $\mathcal{G}^*_U$. The subspace $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ intersects $\mathcal{G}^*_U$ at $\boldsymbol{x}^* = \boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}$, where $\boldsymbol{y}^*$ is infeasible for both (RP) and (RP$\mathcal{X}$).

three $\boldsymbol{x}^*_\top$'s denoted as $\boldsymbol{x}^*_1$, $\boldsymbol{x}^*_2$ and $\boldsymbol{x}^*_3$. These three points are contained in three different G^*'s, denoted as G^*_1, G^*_2 and G^*_3, which are defined in (1.13), (1.14) and (1.15), respectively. Thus, the set of feasible solutions is given by $G = G^*_1 \cup G^*_2 \cup G^*_3$.

2.3 Geometric interpretation of the randomly reduced problems (RP) and (RP$\mathcal{X}$)

We now provide a geometric description of random embeddings in the context of (RP) and (RP$\mathcal{X}$) defined on page 20 for functions with low effective dimensionality.

By applying the ‘random embedding’ (RP) or (RP$\mathcal{X}$), we switch from optimizing over $\mathbb{R}^D$ to optimizing over a lower-dimensional $\mathbb{R}^d$. As illustrated in Figure 2.1, the linear mapping $\boldsymbol{y} \to \boldsymbol{A}\boldsymbol{y} + \boldsymbol{p}$ maps points from $\mathbb{R}^d$ to points along the affine subspace $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ in $\mathbb{R}^D$. Let $\mathcal{G}^*_U$ and $\mathcal{G}^*$ be affine subspaces of global minimizers of (UP) and (P), respectively; for the sake of illustration, assume that $\mathcal{G}^*_U = \mathcal{G}^*$. For (RP)/(RP$\mathcal{X}$) to be successful, it is sufficient that the feasible set of solutions in $\mathbb{R}^d$ after being mapped to $\mathbb{R}^D$ (the line segment between two diamonds for (RP) and the red line segment for (RP$\mathcal{X}$) in Figure 2.1), intersects the set $\mathcal{G}^*_U/G^*$, respectively. Note that, in Figure 2.1, $\mathcal{G}^*_U$ and the feasible set of (RP) do not intersect, which is also the case for G^* and the feasible set of (RP$\mathcal{X}$). Nonetheless, the affine subspace that contains G^* and $\mathcal{G}^*_U$ (that is, $\mathcal{G}^*_U$) and the affine subspace that contains the feasible sets of (RP) and (RP$\mathcal{X}$) (that is, $\boldsymbol{p} + \operatorname{range}(\boldsymbol{A})$) do intersect. This is related to the following theorem in Wang et al. [163]:

Theorem 2.21. *[163, Theorem 2] Let Assumption LowED hold and let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix with $d \geq d_e$. Then, with probability one, for any fixed $\boldsymbol{x} \in \mathbb{R}^D$, there exists a $\boldsymbol{y} \in \mathbb{R}^d$ such that $f(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{y})$. In particular, for a global minimizer $\boldsymbol{x}^*$ of* (UP) *(or* (P)*), with probability one, there exists a $\boldsymbol{y}^* \in \mathbb{R}^d$ such that $f(\boldsymbol{A}\boldsymbol{y}^*) = f(\boldsymbol{x}^*)$.*

Remark 2.22. Theorem 2.21 can be easily extended to affine subspaces. In other words, given the conditions in Theorem 2.21, for a global minimizer $\boldsymbol{x}^*$ and for any $\boldsymbol{p} \in \mathbb{R}^D$, with probability one there exists $\boldsymbol{y}^* \in \mathbb{R}^d$ such that $f(\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}) = f(\boldsymbol{x}^*)$.

Theorem 2.21 establishes that if the dimension of the embedded subspaces (d) is greater than the effective dimension (d_e) of f then there is a global minimizer $\boldsymbol{y}^*$ in $\mathbb{R}^d$ for which the corresponding point $\boldsymbol{x}^*$ in $\mathbb{R}^D$ is a global minimizer. Thus, it is possible to recover a global minimum by just searching in $\mathbb{R}^d$.

Consider now (RP) and (RP$\mathcal{X}$) in light of Theorem 2.21. For $d \geq d_e$, the question of success of (RP) and (RP$\mathcal{X}$) reduces to the question of feasibility of $\boldsymbol{y}^*$ with respect to the corresponding problem, namely, $\boldsymbol{y}^* \in \mathcal{Y}$ for (RP) and $\boldsymbol{A}\boldsymbol{y}^* \in \mathcal{X}$ for (RP$\mathcal{X}$).

In the next section, we characterize some global minimizer(s) $\boldsymbol{y}^*$ of (RP) and (RP$\mathcal{X}$).

2.4 Characterizing minimizers in the reduced space

Before we begin, we make an important disclaimer. The analysis presented in this section can be equally applied to both (RP) and (RP$\mathcal{X}$). To avoid unnecessary re-derivations of the same results for two frameworks separately, we proceed by assuming that $\boldsymbol{x}^*$ is a global minimizer of (UP). The reader should keep in mind that the analysis in this section equally holds for any global minimizer $\boldsymbol{x}^*$ of (P) and the corresponding reduced problem (RP$\mathcal{X}$); in other words, the following results hold if we replace (UP) by (P) and $\mathcal{G}_U^*$ by $\mathcal{G}^*$.

The following theorem provides a useful characterization of global minimizers $\boldsymbol{y}^*$ of (RP)/(RP$\mathcal{X}$). The theorem and its proof were inspired by the proofs of Theorems 2 and 3 in [163].

Theorem 2.23. *Suppose Assumption LowED holds. Let $\boldsymbol{x}^*$ be a global minimizer of* (UP) *with Euclidean projection $\boldsymbol{x}^*_\top$ on the effective subspace, and $\boldsymbol{p} \in \mathbb{R}^D$, a given vector. Let $\mathcal{G}_U^* = \boldsymbol{x}^*_\top + \mathcal{T}^\perp$ and let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix with $d \geq d_e$. Then, $\boldsymbol{y}^* \in \mathbb{R}^d$ satisfies $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} \in \mathcal{G}_U^*$ if and only if*

$$\boldsymbol{B}\boldsymbol{y}^* = \boldsymbol{z}^*, \tag{2.13}$$

where $\boldsymbol{B} = \boldsymbol{U}^T\boldsymbol{A}$ *and where* $\boldsymbol{z}^* \in \mathbb{R}^{d_e}$ *is uniquely defined by*

$$\boldsymbol{U}\boldsymbol{z}^* = \boldsymbol{x}^*_\top - \boldsymbol{p}_\top, \ \textit{with}\ \boldsymbol{p}_\top = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{p}. \tag{2.14}$$

Proof. Let $\boldsymbol{y}^* \in \mathbb{R}^d$ be such that $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} \in \mathcal{G}^*_U$. First, we establish that

$$\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} \in \mathcal{G}^*_U \text{ if and only if } \boldsymbol{x}^*_\top - \boldsymbol{p}_\top = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{y}^*. \tag{2.15}$$

Suppose that $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} \in \mathcal{G}^*_U$. Then, using the definition of $\mathcal{G}^*_U$ we can write $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} = \boldsymbol{x}^*_\top + \tilde{\boldsymbol{x}}$ for some $\tilde{\boldsymbol{x}} \in \mathcal{T}^\perp$. We have

$$\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}_\top = \boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}) = \boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{x}^*_\top + \tilde{\boldsymbol{x}}) = \boldsymbol{x}^*_\top,$$

where we have used $\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}^*_\top = \boldsymbol{x}^*_\top$ and $\boldsymbol{U}\boldsymbol{U}^T\tilde{\boldsymbol{x}} = \boldsymbol{0}$. Conversely, assume that $\boldsymbol{y}^*$ satisfies

$$\boldsymbol{x}^*_\top - \boldsymbol{p}_\top = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{y}^*. \tag{2.16}$$

Denote by $\boldsymbol{S}$ the $D \times D$ orthogonal matrix $(\boldsymbol{U}\ \ \boldsymbol{V})$, where $\boldsymbol{V}$ is defined in Assumption LowED. Using (2.16) and the identity $\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{S}\boldsymbol{S}^T = \boldsymbol{I}_D$, we obtain

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} &= (\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{V}\boldsymbol{V}^T)(\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}) \\
&= \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{p} + \boldsymbol{V}\boldsymbol{V}^T(\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}) \\
&= \boldsymbol{x}^*_\top - \boldsymbol{p}_\top + \boldsymbol{p}_\top + \boldsymbol{V}\boldsymbol{V}^T(\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}) \\
&= \boldsymbol{x}^*_\top + \boldsymbol{V}\boldsymbol{V}^T(\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}).
\end{aligned}$$

Note that $\boldsymbol{V}\boldsymbol{V}^T(\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p})$ lies on $\mathcal{T}^\perp$ as it is the orthogonal projection of $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}$ onto $\mathcal{T}^\perp$, which implies that $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} \in \mathcal{G}^*_U$. This completes the proof of (2.15).

Now we show that (2.13) and (2.16) are equivalent. We multiply both sides of $\boldsymbol{x}^*_\top - \boldsymbol{p}_\top = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{y}^*$ by $\boldsymbol{S}^T$, and obtain

$$\begin{pmatrix}\boldsymbol{U}^T\\ \boldsymbol{V}^T\end{pmatrix}(\boldsymbol{x}^*_\top - \boldsymbol{p}_\top) = \begin{pmatrix}\boldsymbol{U}^T\\ \boldsymbol{V}^T\end{pmatrix}\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{y}^*. \tag{2.17}$$

Since $\boldsymbol{x}^*_\top - \boldsymbol{p}_\top$ is in the column span of $\boldsymbol{U}$, we can write $\boldsymbol{x}^*_\top - \boldsymbol{p}_\top = \boldsymbol{U}\boldsymbol{z}^*$ for some (unique) vector $\boldsymbol{z}^* \in \mathbb{R}^{d_e}$. By substituting the above into (2.17) we obtain

$$\begin{pmatrix}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{z}^*\\ \boldsymbol{V}^T\boldsymbol{U}\boldsymbol{z}^*\end{pmatrix} = \begin{pmatrix}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{y}^*\\ \boldsymbol{V}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{y}^*\end{pmatrix}.$$

This reduces to

$$\begin{pmatrix}\boldsymbol{z}^*\\ \boldsymbol{0}\end{pmatrix} = \begin{pmatrix}\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{y}^*\\ \boldsymbol{0}\end{pmatrix},$$

where we have used the identities $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{V}^T\boldsymbol{U} = \boldsymbol{0}$. To obtain (2.16) from (2.13), multiply (2.13) by $\boldsymbol{U}$. □

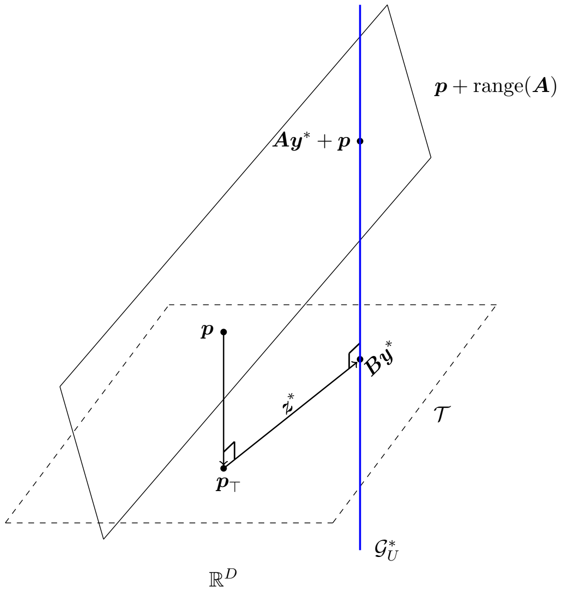

$\boldsymbol{p} + \text{range}(\boldsymbol{A})$
$\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}$
$\boldsymbol{p}$
$\boldsymbol{B}\boldsymbol{y}^*$
$\boldsymbol{z}^*$
$\mathcal{T}$
$\boldsymbol{p}_\top$
$\mathcal{G}_U^*$
$\mathbb{R}^D$

Figure 2.2: Abstract illustration of the result in Theorem 2.23. The intersection point $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}$ between affine subspaces $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ and $\mathcal{G}_U^*$ is projected onto $\mathcal{T}$. The projected point — in the coordinate system of the linear subspace $\mathcal{T}$— is equal to $\boldsymbol{B}\boldsymbol{y}^*$, which is in turn equal to $\boldsymbol{z}^*$ as stated by the theorem. Note that $\boldsymbol{z}^*$ is the vector in $\mathcal{T}$ that connects $\boldsymbol{p}_\top$, the projection point of $\boldsymbol{p}$ onto $\mathcal{T}$, and the point of intersection of $\mathcal{T}$ with $\mathcal{G}_U^*$.

Refer to Figure 2.2 for visualization of the result in Theorem 2.23. We see that the point $\boldsymbol{B}\boldsymbol{y}^*$ (in the coordinate system of $\mathcal{T}$) corresponds to the vector in $\mathbb{R}^D$ that is the projection point of $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}$ onto $\mathcal{T}$. Note also that $\boldsymbol{z}^*$ is the vector in $\mathcal{T}$ that connects $\boldsymbol{p}_\top$, the projection point of $\boldsymbol{p}$ onto $\mathcal{T}$, and the point of intersection of $\mathcal{T}$ with $\mathcal{G}_U^*$. Theorem 2.23 says that $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}$ belongs to $\mathcal{G}_U^*$ if and only if $\boldsymbol{B}\boldsymbol{y}^* = \boldsymbol{z}^*$.

Remark 2.24. Thereafter, we write $\boldsymbol{B}$ to refer to the $d_e \times d$ Gaussian matrix[3] $\boldsymbol{U}^T\boldsymbol{A}$. Furthermore, we write $\boldsymbol{z}^*$ to refer to the $d_e \times 1$ vector that satisfies $\boldsymbol{U}\boldsymbol{z}^* = \boldsymbol{x}_\top^* - \boldsymbol{p}_\top$. Observe that $\|\boldsymbol{z}^*\| = \|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|$ since $\boldsymbol{U}$ is orthogonal. Note also that in the proof of Theorem 2.23 we did not use the fact that $\boldsymbol{A}$ is a Gaussian matrix. In other words, Theorem 2.23 holds for an arbitrary $D \times d$ embedding matrix as long as $d \geq d_e$. Nonetheless, we state Theorem 2.23 with a Gaussian matrix $\boldsymbol{A}$ for consistency and notational convenience.

Theorem 2.23 helps us quantify the number of solutions $\boldsymbol{y}^*$ for which $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}$ lie in the subspace of global minimizers $\mathcal{G}_U^*$. It turns out that there are infinitely many of them if $d > d_e$ and only one if $d = d_e$.

Corollary 2.25. *Suppose Assumption LowED holds. Let* $\boldsymbol{x}^*$ *be a global minimizer of* (UP) *with Euclidean projection* $\boldsymbol{x}_\top^*$ *on the effective subspace, and* $\boldsymbol{p} \in \mathbb{R}^D$*, a given vector. Let*

---

[3] Since $\boldsymbol{U}$ is orthogonal and $\boldsymbol{A}$ is a Gaussian matrix, from Theorem 2.2, it follows that $\boldsymbol{B} = \boldsymbol{U}^T\boldsymbol{A}$ is also a Gaussian matrix.

$\mathcal{G}_U^* = \boldsymbol{x}_\top^* + \mathcal{T}^\perp$ and let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix with $d \geq d_e$. Let $S^* := \{\boldsymbol{y}^* \in \mathbb{R}^d : \boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} \in \mathcal{G}_U^*\}$. Then, the following hold:

- *If $d = d_e$, then S^* has exactly one element with probability 1.*
- *If $d > d_e$, then S^* has infinitely many elements with probability 1.*

Proof. It follows from Theorem 2.23 that the set S^* and the set of solutions to $\boldsymbol{B}\boldsymbol{y} = \boldsymbol{z}^*$ coincide. According to Theorem 2.4, $\boldsymbol{B}\boldsymbol{B}^T$ is positive definite with probability 1, which implies that $\text{rank}(\boldsymbol{B}\boldsymbol{B}^T) = d_e$ with probability 1. Since $\text{rank}(\boldsymbol{B}) = \text{rank}(\boldsymbol{B}\boldsymbol{B}^T)$, $\text{rank}(\boldsymbol{B}) = d_e$ with probability 1. Hence, the linear system $\boldsymbol{B}\boldsymbol{y} = \boldsymbol{z}^*$ almost surely has a solution. If $d = d_e$ the linear system has only one solution. If $d > d_e$ the system is underdetermined and, therefore, has infinitely many solutions. □

For (RP) and (RP$\mathcal{X}$) to be successful, it is sufficient that one of the minimizers $\boldsymbol{y}^*$ is feasible for the respective reduced problems. We proceed further by choosing one particular minimizer that is easy to analyse and that is close to the origin in some norm so that the likelihood of falling inside the feasible sets of (RP) and (RP$\mathcal{X}$) is relatively high. An obvious candidate is the minimal Euclidean norm solution,

$$\begin{aligned} \boldsymbol{y}_2^* = \underset{\boldsymbol{y} \in \mathbb{R}^d}{\text{argmin}} \ & \|\boldsymbol{y}\|_2 \\ \text{s.t.} \ & \boldsymbol{y} \in \mathcal{S}^*, \end{aligned} \tag{2.18}$$

where S^* is defined[4] in Corollary 2.25. It is straightforward to derive the closed form solution of $\boldsymbol{y}_2^*$.

Corollary 2.26. *Suppose Assumption LowED holds. Let $\boldsymbol{x}^*$ be a global minimizer of* (UP) *with Euclidean projection $\boldsymbol{x}_\top^*$ on the effective subspace, and $\boldsymbol{p} \in \mathbb{R}^D$, a given vector. Let $\mathcal{G}_U^* = \boldsymbol{x}_\top^* + \mathcal{T}^\perp$ and let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix with $d \geq d_e$. Let $\boldsymbol{y}_2^*$ be defined as in* (2.18)*. Then,*

$$\boldsymbol{y}_2^* = \boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{z}^*, \tag{2.19}$$

where $\boldsymbol{z}^*$ is defined in (2.14)*.*

Proof. It follows from Theorem 2.23 that $\boldsymbol{y}_2^*$ must be the solution of the following problem

$$\begin{aligned} \min \ & \|\boldsymbol{y}\|_2^2 \\ \text{s.t.} \ & \boldsymbol{B}\boldsymbol{y} = \boldsymbol{z}^*, \end{aligned}$$

which has the solution (2.19). □

In what follows, we derive the distribution of $\boldsymbol{y}_2^*$, which we rely upon in Chapters 3 and 4 to estimate the probabilities of success of (RP) and (RP$\mathcal{X}$).

---

[4] Note that in the context of the constrained problem (P) and the reduced problem (RP$\mathcal{X}$), S^* will be defined through $\mathcal{G}^*$, that is, $S^* := \{\boldsymbol{y}^* \in \mathbb{R}^d : \boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} \in \mathcal{G}^*\}$.

2.4.1 Distributional characterization of the minimal Euclidean norm minimizer

In this section, we derive the distribution of the Euclidean norm of $\boldsymbol{y}_2^*$ and show that $\boldsymbol{y}_2^*$ follows a spherical distribution. Then, using these two facts, we derive the p.d.f. of $\boldsymbol{y}_2^*$. The following theorem establishes a notable relationship between $\|\boldsymbol{y}_2^*\|$ and the chi-squared distribution.

Theorem 2.27. *Suppose Assumption LowED holds. Let $\boldsymbol{x}^*$ be a global minimizer of* (UP) *and $\boldsymbol{p} \in \mathbb{R}^D$ a given vector, with respective projections $\boldsymbol{x}_\top^*$ and $\boldsymbol{p}_\top$ on the effective subspace $\mathcal{T}$. Let $\mathcal{G}_U^* = \boldsymbol{x}_\top^* + \mathcal{T}^\perp$ and let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix with $d \geq d_e$. Let $\boldsymbol{y}_2^*$ be defined in* (2.18)*. Then, $\boldsymbol{y}_2^*$ satisfies*

$$\frac{\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|_2^2}{\|\boldsymbol{y}_2^*\|_2^2} \sim \chi^2_{d-d_e+1} \quad \text{if } \boldsymbol{x}_\top^* \neq \boldsymbol{p}_\top. \tag{2.20}$$

If $\boldsymbol{x}_\top^* = \boldsymbol{p}_\top$, then $\boldsymbol{y}_2^* = \boldsymbol{0}$.

Proof. For $\boldsymbol{y}_2^*$, from (2.19), we have

$$\begin{aligned} \|\boldsymbol{y}_2^*\|_2^2 &= (\boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{z}^*)^T\boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{z}^* \\ &= \boldsymbol{z}^{*T}(\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{z}^*, \end{aligned} \tag{2.21}$$

Assume that $\boldsymbol{x}_\top^* \neq \boldsymbol{p}_\top$. Using the fact $\|\boldsymbol{z}^*\| = \|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|$ and Lemma 2.6, we obtain (2.20):

$$\frac{\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|_2^2}{\|\boldsymbol{y}_2\|_2^2} = \frac{\|\boldsymbol{z}^*\|_2^2}{\|\boldsymbol{y}_2\|_2^2} = \frac{\|\boldsymbol{z}^*\|^2}{\boldsymbol{z}^{*T}(\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{z}^*} \sim \chi^2_{d-d_e+1}.$$

If $\boldsymbol{x}_\top^* = \boldsymbol{p}_\top$, then $\boldsymbol{z}^* = \boldsymbol{0}$ and $\boldsymbol{y}_2^* = \boldsymbol{0}$ by (2.21). ☐

The above result is equivalent to saying that $\|\boldsymbol{y}_2^*\|^2/\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|^2$ follows the inverse chi-squared distribution with $d - d_e + 1$ degrees of freedom (see Definition 2.8). Theorem 2.27 reveal a linear dependence of $\|\boldsymbol{y}_2^*\|$ on $\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|$; larger values of $\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|$ contribute to the increase in the likelihood of $\boldsymbol{y}_2^*$ being further away from the origin. The results also imply that $\|\boldsymbol{y}_2^*\|$ is independent of D as long as $\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|$ is fixed.

Corollary 2.28. *Suppose Assumption LowED holds. Let $\boldsymbol{x}^*$ be a global minimizer of* (UP) *and $\boldsymbol{p} \in \mathbb{R}^D$ a given vector, with respective projections $\boldsymbol{x}_\top^*$ and $\boldsymbol{p}_\top$ on the effective subspace $\mathcal{T}$. Let $\mathcal{G}_U^* = \boldsymbol{x}_\top^* + \mathcal{T}^\perp$ and let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix with $d \geq d_e$. Let $\boldsymbol{y}_2^*$ be defined in* (2.18)*. Provided that $d - d_e > 1$ we have*

$$\mathbb{E}[\|\boldsymbol{y}_2^*\|_2^2] = \frac{\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|_2^2}{d - d_e - 1}. \tag{2.22}$$

Proof. Let $W \sim 1/\chi^2_{d-d_e+1}$. Then, $W \overset{law}{=} \|\boldsymbol{y}_2^*\|_2^2/\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|_2^2$ by Theorem 2.27 if $\boldsymbol{x}_\top^* \neq \boldsymbol{p}_\top$. By applying Lemma 2.9, we obtain

$$\mathbb{E}[\|\boldsymbol{y}_2^*\|_2^2] = \mathbb{E}[\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|_2^2 W] = \frac{\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|_2^2}{d - d_e - 1}$$

for $d - d_e > 1$. If $\boldsymbol{x}_\top^* = \boldsymbol{p}_\top$ then $\boldsymbol{y}_2^* = \boldsymbol{0}$ and (2.22) follows immediately. ☐

The expected value in (2.22) is inversely proportional to $d - d_e$. In other words, for a fixed value of d_e, larger values of the dimension of the embedding subspace bring $\boldsymbol{y}_2^*$ closer to the origin.

Unsurprisingly, it is possible to derive the p.d.f. of a spherically distributed random vector (see Definition 2.13) while only knowing the distribution of its Euclidean norm. Next, we show that $\boldsymbol{y}_2^*$ has a spherical distribution and derive its p.d.f. using Theorem 2.27.

Theorem 2.29 ($\boldsymbol{y}_2^*$ follows a spherical distribution). *Suppose Assumption LowED holds. Let $\boldsymbol{x}^*$ be a global minimizer of* (UP) *and $\boldsymbol{p} \in \mathbb{R}^D$ a given vector, with respective projections $\boldsymbol{x}_\top^*$ and $\boldsymbol{p}_\top$ on the effective subspace $\mathcal{T}$. Assume that $\boldsymbol{x}_\top^* \neq \boldsymbol{p}_\top$. Let $\mathcal{G}_U^* = \boldsymbol{x}_\top^* + \mathcal{T}^\perp$ and let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix with $d \geq d_e$. Let $\boldsymbol{y}_2^*$ be defined in* (2.18)*. Then, $\boldsymbol{y}_2^*$ follows a spherical distribution.*

Proof. Let $\boldsymbol{S}$ be any $d \times d$ orthogonal matrix. Let $h : \mathbb{R}^{d_e d \times 1} \to \mathbb{R}^{d \times 1}$ be a vector-valued function defined as

$$h(\text{vec}(\boldsymbol{B})) = \boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{z}^*,$$

where vec($\boldsymbol{B}$) denotes the $Dd \times 1$ vector $(\boldsymbol{b}_1^T \; \boldsymbol{b}_2^T \; \cdots \boldsymbol{b}_d^T)^T$ with $\boldsymbol{b}_i$ being the ith column vector of $\boldsymbol{B}$. Using the fact that the inverse of a matrix is equal to the ratio of its adjugate to its determinant we can express h as

$$h(\text{vec}(\boldsymbol{B})) = \begin{pmatrix} \frac{p_1(\boldsymbol{B})}{q(\boldsymbol{B})} & \frac{p_2(\boldsymbol{B})}{q(\boldsymbol{B})} & \cdots & \frac{p_d(\boldsymbol{B})}{q(\boldsymbol{B})} \end{pmatrix}^T,$$

where $p_i(\boldsymbol{B})$ for $1 \leq i \leq d$ are some polynomials in the entries of $\boldsymbol{B}$ and $q(\boldsymbol{B})$ is the determinant of $\boldsymbol{B}\boldsymbol{B}^T$. Since q and p_i's are polynomials in Gaussian random variables, they are all measurable. Furthermore, since $\boldsymbol{B}$ is Gaussian, by Theorem 2.4, $\mathbb{P}[q = 0] = 0$; this implies that p_i/q is a measurable function for each $i = 1, 2, \ldots, d$ (see [165, Theorem 4.10]).

For $\boldsymbol{y}_2^* = \boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{z}^*$, we have

$$\boldsymbol{y}_2^* = h(\text{vec}(\boldsymbol{B})) \text{ and } \boldsymbol{S}\boldsymbol{y}_2^* = h(\text{vec}(\boldsymbol{B}\boldsymbol{S}^T)).$$

According to Theorem 2.2, $\text{vec}(\boldsymbol{B}) \overset{law}{=} \text{vec}(\boldsymbol{B}\boldsymbol{S}^T)$. Then, by applying Lemma A.7, we obtain

$$\boldsymbol{y}_2^* = h(\text{vec}(\boldsymbol{B})) \overset{law}{=} h(\text{vec}(\boldsymbol{B}\boldsymbol{S}^T)) = \boldsymbol{S}\boldsymbol{y}_2^*.$$

Hence, $\boldsymbol{y}_2^*$ follows a spherical distribution by Definition 2.13. ☐

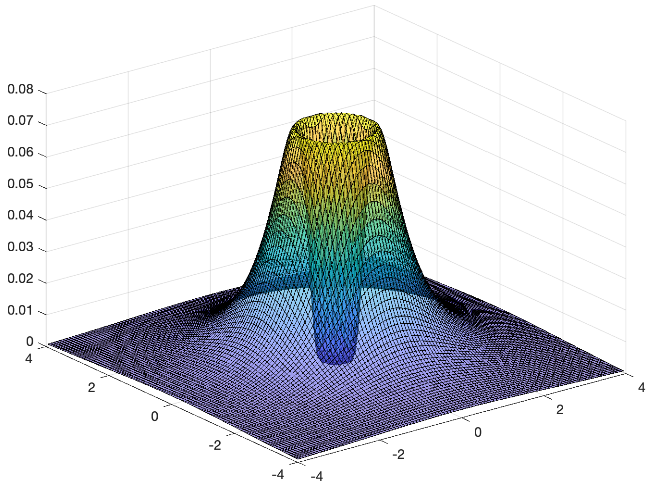


Figure 2.3: An illustration of the p.d.f. of $\boldsymbol{y}_2^*$ for $d = 2$, $n = 2$ and $\boldsymbol{x}_\top^* = [1\ 1]^T$, $\boldsymbol{p} = \boldsymbol{0}$.

We are now ready to derive the p.d.f. of $\boldsymbol{y}_2^*$.

Theorem 2.30 (The p.d.f. of $\boldsymbol{y}_2^*$). *Suppose Assumption LowED holds. Let* $\boldsymbol{x}^*$ *be a global minimizer of* (UP) *and* $\boldsymbol{p} \in \mathbb{R}^D$ *a given vector, with respective projections* $\boldsymbol{x}_\top^*$ *and* $\boldsymbol{p}_\top$ *on the effective subspace. Assume that* $\boldsymbol{x}_\top^* \neq \boldsymbol{p}_\top$. *Let* $\mathcal{G}_U^* = \boldsymbol{x}_\top^* + \mathcal{T}^\perp$ *and let* $\boldsymbol{A}$ *be a* $D \times d$ *Gaussian matrix with* $d \geq d_e$. *Let* $\boldsymbol{y}_2^*$ *be defined in* (2.18). *Then, the probability density function of* $\boldsymbol{y}_2^*$ *is given by*

$$g^*(\boldsymbol{y}) = \pi^{-d/2}\left(\frac{\Gamma(d/2)}{\Gamma(n/2)}\right)\left(\frac{\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|}{\sqrt{2}}\right)^n (\boldsymbol{y}^T\boldsymbol{y})^{-(n+d)/2} e^{-\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|^2/(2\boldsymbol{y}^T\boldsymbol{y})},$$

where $n = d - d_e + 1$.

Proof. To simplify the derivations, let us assume for now that $\|\boldsymbol{z}^*\| = 1$. Let us first show that $\mathbb{P}[\boldsymbol{y}_2^* = \boldsymbol{0}] = 0$. Let $W \sim 1/\chi^2_{d-d_e+1}$. We have

$$\mathbb{P}[\boldsymbol{y}_2^* = \boldsymbol{0}] = \mathbb{P}[\|\boldsymbol{y}_2^*\|^2 = 0] = \mathbb{P}[W = 0],$$

where the last equality follows from Theorem 2.27. Since W is a continuous random variable, $\mathbb{P}[W = 0] = 0$; hence, $\mathbb{P}[\boldsymbol{y}_2^* = \boldsymbol{0}] = 0$. Given this fact and the fact that $\boldsymbol{y}_2^*$ follows a spherical distribution (Theorem 2.29), Theorem 2.15 allows us to express the p.d.f. $g(\cdot)$ of $\boldsymbol{y}_2^*$ in terms of the p.d.f. $h(\cdot)$ of $\|\boldsymbol{y}_2^*\|$, which we derive in Lemma A.6. Theorem 2.15 gives

$$g(\hat{\boldsymbol{y}}) = \frac{\Gamma(d/2)}{2\pi^{d/2}}(\hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}})^{(1-d)/2} h(\|\boldsymbol{y}_2^*\|). \tag{2.23}$$

By using (A.4) for $h(\cdot)$ in (2.23), we obtain

$$g(\hat{\boldsymbol{y}}) = 2^{-n/2}\pi^{-d/2}\frac{\Gamma(d/2)}{\Gamma(n/2)}(\hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}})^{-(n+d)/2}e^{-1/(2\hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}})}. \tag{2.24}$$

To derive the p.d.f. for an arbitrary non-zero $\boldsymbol{z}^*$, we consider the linear transformation $\bar{\boldsymbol{y}} = \|\boldsymbol{z}^*\|\hat{\boldsymbol{y}}$. The Jacobian of the transformation is equal to $1/\|\boldsymbol{z}^*\|^d$. Thus, the p.d.f. $\bar{g}(\bar{\boldsymbol{y}})$ of $\bar{\boldsymbol{y}}$ satisfies

$$\bar{g}(\bar{\boldsymbol{y}}) = \frac{g(\bar{\boldsymbol{y}}/\|\boldsymbol{z}^*\|)}{\|\boldsymbol{z}\|^d},$$

which together with (2.24) and the fact $\|\boldsymbol{z}^*\| = \|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|$ yields the desired result. $\square$

Figure 2.3 illustrates the p.d.f. of a two-dimensional $\boldsymbol{y}_2^*$. The shape of the p.d.f. resembles a volcano with the mass concentrated at a certain distance from the origin suggesting that $\boldsymbol{y}_2^*$ is unlikely to be neither too close to, nor too distant from the origin. We also note that the p.d.f. has no direct dependence on D.

Chapter 3

Unconstrained optimization of functions with low effective dimensionality

In this chapter, we estimate the probability of success of (RP) based on the results we obtained for the distribution of the random minimizer $\boldsymbol{y}_2^*$ defined in (2.18). We then test the effectiveness of (RP) numerically to validate our theoretical investigations.

3.1 Bounding the success of the reduced problem (RP)

We lower bound the probability of success of (RP) using Theorem 2.27 and its consequences. To derive the bound, we make use of the following corollary implied by Theorem 2.27.

Corollary 3.1. *Suppose Assumption LowED holds. Let* $\boldsymbol{x}^*$ *be a global minimizer of* (UP) *and* $\boldsymbol{p} \in \mathcal{X}$ *a given vector, with respective projections* $\boldsymbol{x}_\top^*$ *and* $\boldsymbol{p}_\top$ *on the effective subspace. Assume that* $\boldsymbol{x}_\top^* \neq \boldsymbol{p}_\top$*. Let* $\boldsymbol{A}$ *be a* $D \times d$ *Gaussian matrix with* $d \geq d_e$*. Let* $\boldsymbol{y}_2^*$ *be defined in* (2.18)*. Then,*

$$\mathbb{P}[\|\boldsymbol{y}_2^*\|_2 \leq \delta] = \mathbb{P}\left[\chi^2_{d-d_e+1} \geq \frac{\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|_2^2}{\delta^2}\right]$$

for any $\delta > 0$.

Proof. For any $\epsilon > 0$, we have

$$\mathbb{P}\left[\|\boldsymbol{y}_2^*\|_2 \leq \frac{\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|_2}{\epsilon}\right] = \mathbb{P}\left[\frac{\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|_2^2}{\|\boldsymbol{y}_2^*\|_2^2} \geq \epsilon^2\right] = \mathbb{P}[\chi^2_{d-d_e+1} \geq \epsilon^2],$$

where the second equality follows from Theorem 2.27. By letting $\epsilon = \|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|_2/\delta$, we obtain the desired result. □

Remark 3.2. Note that in Corollary 3.1 we assume that $\boldsymbol{x}_\top^* \neq \boldsymbol{p}_\top$. If $\boldsymbol{x}_\top^* = \boldsymbol{p}_\top$, then $f(\boldsymbol{p}) = f(\boldsymbol{p}_\top + \boldsymbol{p}_\perp) = f(\boldsymbol{x}_\top^* + \boldsymbol{p}_\perp) = f_U^*$ implying that $\boldsymbol{p}$ is a global minimizer so that, for any embedding, (RP) is successful with a trivial solution $\boldsymbol{y}^* = \boldsymbol{0}$.

Using Corollary 3.1 we now prove the main result of this chapter.

Theorem 3.3. *Let Assumption LowED hold. Let* $\boldsymbol{p} \in \mathbb{R}^D$ *be a given vector and let* $\boldsymbol{A}$ *be a* $D \times d$ *Gaussian matrix with* $d \geq d_e$*. Then, for any* $\delta > 0$*, we have*

$$\mathbb{P}[(\mathrm{RP}) \textit{ is successful}] \geq \mathbb{P}\left[\chi^2_{d-d_e+1} \geq \frac{\min_{\boldsymbol{x}^* \in \mathcal{G}_U} \|\boldsymbol{x}^* - \boldsymbol{p}\|_2^2}{\delta^2}\right], \tag{3.1}$$

where $\mathcal{G}_U$ *is defined in* (2.9)*.*

Proof. Note the following relationship between the probabilities:

$$\mathbb{P}[(\mathrm{RP}) \text{ is successful}] \geq \mathbb{P}[\boldsymbol{y}_2^* \in \mathcal{Y}] = \mathbb{P}[\|\boldsymbol{y}_2^*\|_\infty \leq \delta] \geq \mathbb{P}[\|\boldsymbol{y}_2^*\|_2 \leq \delta], \tag{3.2}$$

where the first inequality follows from Definition 1.3 and the definition of $\boldsymbol{y}_2^*$, and where the second inequality is implied by $\|\boldsymbol{y}_2^*\|_\infty \leq \|\boldsymbol{y}_2^*\|_2$. By applying Corollary 3.1 to the last probability in (3.2), we obtain

$$\mathbb{P}[(\mathrm{RP}) \text{ is successful}] \geq \mathbb{P}\left[\chi^2_{d-d_e+1} \geq \frac{\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|_2^2}{\delta^2}\right] = \mathbb{P}\left[\chi^2_{d-d_e+1} \geq \frac{\|\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{x}^* - \boldsymbol{p})\|_2^2}{\delta^2}\right] \tag{3.3}$$

for any $\delta > 0$ and any $\boldsymbol{x}^* \in \mathcal{G}_U$ such that $\|\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{x}^* - \boldsymbol{p})\|_2 \neq 0$, where $\mathcal{G}_U$ is defined in (2.9). Note that (3.3) also holds for $\boldsymbol{x}^*$ with $\|\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{x}^* - \boldsymbol{p})\|_2 = 0$ since, in this case, (RP) is successful with probability 1 (see Remark 3.2). Hence, (3.3) holds for any $\boldsymbol{x}^* \in \mathcal{G}_U$, which then implies

$$\begin{aligned}\mathbb{P}[(\mathrm{RP}) \text{ is successful}] &\geq \max_{\boldsymbol{x}^* \in \mathcal{G}_U} \mathbb{P}\left[\chi^2_{d-d_e+1} \geq \frac{\|\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{x}^* - \boldsymbol{p})\|_2^2}{\delta^2}\right] \\ &= \mathbb{P}\left[\chi^2_{d-d_e+1} \geq \frac{\min_{\boldsymbol{x}^* \in \mathcal{G}_U} \|\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{x}^* - \boldsymbol{p})\|_2^2}{\delta^2}\right],\end{aligned}$$

where the equality follows from the fact that the tail distribution $\mathbb{P}[X > x]$ of any random variable X is a monotonically decreasing function in x.

In what follows, we show that $\min_{\boldsymbol{x}^* \in \mathcal{G}_U} \|\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{x}^* - \boldsymbol{p})\|_2^2 = \min_{\boldsymbol{x}^* \in \mathcal{G}_U} \|\boldsymbol{x}^* - \boldsymbol{p}\|_2^2$. Define sets $\mathcal{Z} := \{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{U}\boldsymbol{z} = \boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{x}^* - \boldsymbol{p}), \boldsymbol{x}^* \in \mathcal{G}_U\}$ and $\mathcal{S} := \{\boldsymbol{U}\boldsymbol{z} + \boldsymbol{V}\boldsymbol{c} : \boldsymbol{z} \in \mathcal{Z}, \boldsymbol{c} \in \mathbb{R}^{D-d_e}\}$, where $\boldsymbol{V}$ is defined in Assumption LowED. Define also $\mathcal{G}_U(-\boldsymbol{p}) := -\boldsymbol{p} + \mathcal{G}_U$. First, we establish that $\mathcal{G}_U(-\boldsymbol{p}) = \mathcal{S}$ by showing that $\mathcal{G}_U(-\boldsymbol{p}) \subseteq \mathcal{S}$ and that $\mathcal{S} \subseteq \mathcal{G}_U(-\boldsymbol{p})$.

Let $\bar{\boldsymbol{x}}^* \in \mathcal{G}_U(-\boldsymbol{p})$. Then, $\bar{\boldsymbol{x}}^* = \boldsymbol{x}^* - \boldsymbol{p}$ for some $\boldsymbol{x}^* \in \mathcal{G}_U$. We can write $\bar{\boldsymbol{x}}^* = \boldsymbol{x}^* - \boldsymbol{p} = \boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{x}^* - \boldsymbol{p}) + \boldsymbol{V}\boldsymbol{V}^T(\boldsymbol{x}^* - \boldsymbol{p})$ since $\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}$. Let $\boldsymbol{z} = \boldsymbol{U}^T(\boldsymbol{x}^* - \boldsymbol{p})$ and $\boldsymbol{c} = \boldsymbol{V}^T(\boldsymbol{x}^* - \boldsymbol{p})$ and note that $\boldsymbol{z} \in \mathcal{Z}$ and $\boldsymbol{c} \in \mathbb{R}^{D-d_e}$. Hence, $\bar{\boldsymbol{x}}^* \in \mathcal{S}$, which proves that $\mathcal{G}_U(-\boldsymbol{p}) \subseteq \mathcal{S}$.

Let $\bar{\boldsymbol{x}}^* \in \mathcal{S}$. Then, $\bar{\boldsymbol{x}}^* = \boldsymbol{U}\boldsymbol{z} + \boldsymbol{V}\boldsymbol{c}$ for some $\boldsymbol{z} \in \mathcal{Z}$ and $\boldsymbol{c} \in \mathbb{R}^{D-d_e}$. By writing $\boldsymbol{x}^* = \boldsymbol{p} + \bar{\boldsymbol{x}}^*$, we obtain

$$\begin{aligned} f(\boldsymbol{x}^*) = f(\boldsymbol{p} + \bar{\boldsymbol{x}}^*) &= f(\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{p} + \boldsymbol{V}\boldsymbol{V}^T\boldsymbol{p} + \boldsymbol{U}\boldsymbol{z} + \boldsymbol{V}\boldsymbol{c}) \\ &= f(\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{p} + \boldsymbol{U}\boldsymbol{z}) \\ &= f(\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{p} + \boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{x}^* - \boldsymbol{p})) \\ &= f(\boldsymbol{x}^*_\top) = f^*_U, \end{aligned}$$

where the third equality follows from the assumption that f has low effective dimensionality and the fact that $\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{p} + \boldsymbol{V}\boldsymbol{c} \in \mathcal{T}^\perp$, the fourth equality is by definition of $\mathcal{Z}$, the fifth equality follows from the definition of $\boldsymbol{x}^*_\top$ (given in Assumption LowED) and the last equality follows from (2.10). Hence, by definition of $\mathcal{G}_U$, $\boldsymbol{x}^* \in \mathcal{G}_U$ and therefore $\bar{\boldsymbol{x}}^* \in \mathcal{G}_U(-\boldsymbol{p})$. This proves that $\mathcal{S} \subseteq \mathcal{G}_U(-\boldsymbol{p})$.

Finally, we have

$$\begin{aligned} \min_{\boldsymbol{x}^* \in \mathcal{G}_U} \|\boldsymbol{x}^* - \boldsymbol{p}\|_2^2 &= \min_{\bar{\boldsymbol{x}}^* \in \mathcal{G}_U(-\boldsymbol{p})} \|\bar{\boldsymbol{x}}^*\|_2^2 && \text{(by definition of } \mathcal{G}_U(-\boldsymbol{p})) \\ &= \min_{\boldsymbol{z} \in \mathcal{Z},\, \boldsymbol{c} \in \mathbb{R}^{D-d_e}} \|\boldsymbol{U}\boldsymbol{z} + \boldsymbol{V}\boldsymbol{c}\|_2^2 && \text{(since } \mathcal{G}_U(-\boldsymbol{p}) = \mathcal{S} \text{ and by definition of } \mathcal{S}) \\ &= \min_{\boldsymbol{z} \in \mathcal{Z},\, \boldsymbol{c} \in \mathbb{R}^{D-d_e}} \|\boldsymbol{U}\boldsymbol{z}\|_2^2 + \|\boldsymbol{V}\boldsymbol{c}\|_2^2 && \text{(since } \boldsymbol{U}^T\boldsymbol{V} = \boldsymbol{0}, \boldsymbol{V}^T\boldsymbol{U} = \boldsymbol{0}) \\ &= \min_{\boldsymbol{z} \in \mathcal{Z},\, \boldsymbol{c} \in \mathbb{R}^{D-d_e}} \|\boldsymbol{U}\boldsymbol{z}\|_2^2 + \|\boldsymbol{c}\|_2^2 && \text{(since } \boldsymbol{V} \text{ is orthogonal)} \\ &= \min_{\boldsymbol{z} \in \mathcal{Z}} \|\boldsymbol{U}\boldsymbol{z}\|_2^2 + \min_{\boldsymbol{c} \in \mathbb{R}^{D-d_e}} \|\boldsymbol{c}\|_2^2 \\ &= \min_{\boldsymbol{z} \in \mathcal{Z}} \|\boldsymbol{U}\boldsymbol{z}\|_2^2 + 0 \\ &= \min_{\boldsymbol{x}^* \in \mathcal{G}_U} \|\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{x}^* - \boldsymbol{p})\|_2^2 && \text{(by definition of } \mathcal{Z}) \end{aligned}$$

□

Using Theorem 3.3, one can now derive quantifiable bounds for the success of (RP) by applying any tail bound on the chi-squared distribution. We use the bound derived in Lemma A.5.

Corollary 3.4. *Let Assumption LowED hold. Let $\boldsymbol{p} \in \mathbb{R}^D$ be a given vector and let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix with $d \geq d_e$. Let $\mu = \min_{\boldsymbol{x}^* \in \mathcal{G}_U} \|\boldsymbol{x}^* - \boldsymbol{p}\|_2$. Then, for any $\delta > 0$, we have*

$$\mathbb{P}[(\text{RP}) \textit{ is successful}] \geq 1 - C(n)\left(1 + \frac{n}{2}e^{-\mu^2/(2\delta^2)}\right)\left(\frac{\mu}{\sqrt{2}\delta}\right)^n, \tag{3.4}$$

where $n = d - d_e + 1$ and

$$C(n) = \frac{4}{n(n+2)\Gamma\big(n/2\big)}.$$

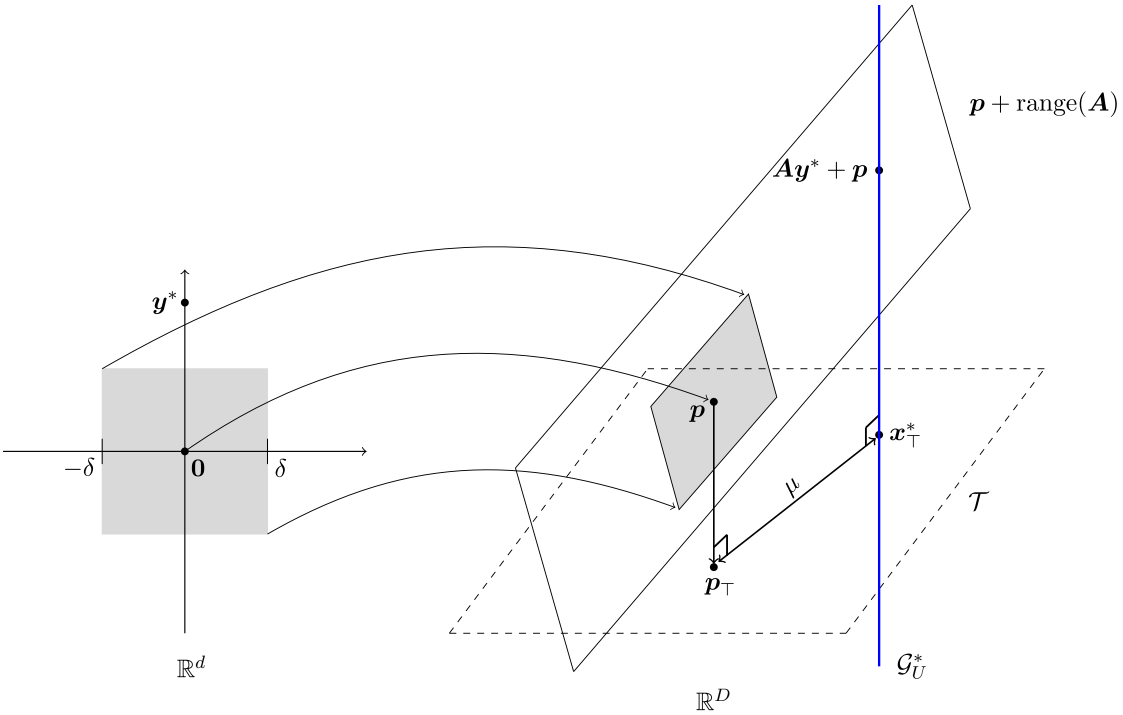


Figure 3.1: Abstract illustration of the random embedding technique similar to Figures 2.1 and 2.2 but with the parameter μ displayed. The size of the $\mathcal{Y} = [-\delta, \delta]^d$ box has a linear relationship with the size of the shaded quadrilateral region (that lies on the embedded subspace $\boldsymbol{p} + \mathrm{range}(\boldsymbol{A})$). This quadrilateral region, in turn, has a linear relationship with μ, that is, scaling of μ by a factor of k requires scaling of the quadrilateral region by the same factor for $\mathbb{P}[\text{(RP) is successful}]$ to remain constant.

Proof. Lemma A.5 implies that

$$\mathbb{P}[\chi^2_n \geq \epsilon^2] \geq 1 - C(n)\left(1 + \frac{n}{2}e^{-\epsilon^2/2}\right)(\epsilon^2/2)^{n/2} \tag{3.5}$$

for any $\epsilon > 0$. By letting $\epsilon = \mu/\delta$ and applying (3.5) to (3.1), we obtain the desired bound. □

Let R^* denote the right hand side of (3.4). First, we note that R^* is a function of μ/δ and $d - d_e$. The bound reveals a linear relationship between μ and δ (refer to Figure 3.1 for visualization of the parameters μ and δ); scaling μ and δ by the same factor does not affect the value of R^*. Furthermore, observe that for smaller values of μ and/or larger values of δ, R^* is closer to 1. Numerical experiments show that for large values of n and/or μ/δ, the bound (3.4) is less tight; this is also signified by the asymptotic behaviour of R^*, when $R^* \to -\infty$ monotonically as $\mu/\delta \to \infty$, making the bound useless for large enough μ/δ.

It is remarkable that R^* has no explicit dependence on D, the dimension of the original optimization problem. This implies that larger D does not diminish the success of the reduced problem as long as μ and d_e are fixed. Dependence of R^* on $d - d_e$ indicates that the success is determined by the value of d relative to d_e and not so much by the individual values of d and d_e. Larger (smaller) values of d with respect to d_e require smaller (larger) δ if R^* is kept constant; knowing this fact is crucial when initializing values of d and δ in practice. It displays a convenient interplay between d and δ allowing more flexibility in choosing one with respect to the other.

Previous bounds. One can derive similar bounds for the success of (RP) by bounding $\mathbb{P}[\|\boldsymbol{y}_2^*\| \leq \delta]$ in (3.2) using the Cauchy-Schwarz inequality. Since $\boldsymbol{y}_2^* = \boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{z}^*$, we have

$$\|\boldsymbol{y}_2^*\| \leq \|\boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}\| \cdot \|\boldsymbol{z}^*\|. \tag{3.6}$$

By using the fact that $\|\boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}\| = 1/s_{\min}(\boldsymbol{B}^T)$, where $s_{\min}(\boldsymbol{B}^T)$ denotes the smallest singular value of $\boldsymbol{B}^T$, we obtain

$$\mathbb{P}[\|\boldsymbol{y}_2^*\| \leq \delta] \geq \mathbb{P}\left[\frac{\|\boldsymbol{z}^*\|}{s_{\min}(\boldsymbol{B}^T)} \leq \delta\right].$$

We can now use any suitable tail bound for the smallest singular value of the Gaussian matrix to bound the latter probability.

Wang et al. [163], by applying the above technique and Theorem A.1 to bound the singular value, derived the following bound

$$\mathbb{P}[\text{(RP) is successful}] \geq 1 - \frac{\mu\sqrt{d_e}}{\delta}.$$

Their derivation is predicated on the assumptions that $d = d_e$ and that $\mathcal{T}$ is spanned by the standard basis vectors. In [141], Sanyang and Kabán extended Wang et al. [163]'s bound allowing a wider range of values for d. Using the bound in Theorem A.2 for $s_{\min}(\boldsymbol{B}^T)$ they showed that

$$\mathbb{P}[\text{(RP) is successful}] \geq 1 - e^{-(\sqrt{d}-\sqrt{d_e}-\mu/\delta)^2/2}$$

assuming certain conditions on d and/or δ are satisfied. One can also use Rudelson and Vershynin's bound (Theorem A.3) to obtain

$$\mathbb{P}[\text{(RP) is successful}] \geq 1 - \left(\frac{C\mu}{\delta(\sqrt{d}-\sqrt{d_e-1})}\right)^{d-d_e+1} - e^{-cd},$$

where $C, c > 0$ are absolute constants. This bound shows dependence of the probability on the difference $d - d_e$, which is also manifest in our bound. The Rudelson and Vershynin's bound cannot be used for practical purposes due to the unknown C and c; we require explicit bounds to define the size of $\mathcal{Y}$.

Unlike the bounds of Wang et al. [163] and Sanyang and Kabán [141], Corollary 3.1 is applicable to any $d \geq d_e$ and an arbitrary subspace $\mathcal{T}$. Moreover, using the exact distribution of $\|\boldsymbol{y}_2^*\|$ given in Corollary 3.1, we circumvent the application of the intermediate Cauchy-Schwarz as in (3.6) and bound the distribution of $\|\boldsymbol{y}_2^*\|$ directly.

3.2 Choices of (RP) parameters

The present section aims to test numerically the quality of the bound (3.4). We will also use the results of this section to select suitable pairs of parameters d and δ for (RP) in the numerical experiments later.

Suppose that we are given a function f satisfying Assumption LowED with the set of global minimizers $\mathcal{G}_U$ consisting of only one $(D - d_e)$-dimensional simply connected $\mathcal{G}_U^*$. Let $\boldsymbol{x}_\top^*$ be defined as in Definition 2.20, $\boldsymbol{z}$ be defined by the equation $\boldsymbol{U}\boldsymbol{z} = \boldsymbol{x}_\top^*$ and $\boldsymbol{p} = \boldsymbol{0}$. We also define $\mu := \min_{\boldsymbol{x}^* \in \mathcal{G}_U} \|\boldsymbol{x}^*\|$ and note that $\mu = \|\boldsymbol{x}_\top^*\|$.

We test (3.4) by contrasting the left-hand side of (3.4) (denoted by L^*) to its right-hand side (denoted by R^*). We compare L^* and R^* for four different values of $d - d_e$, namely, 0, 1, 2 and 3. For each value of $d - d_e$, we express R^* as a function of $\bar{\delta} := \delta/\mu$ and using its closed form we plot R^* for $\bar{\delta} \in [0.02, 10]$. We do not have a closed form expression for L^*, but we can approximate it numerically. In what follows, we describe how this is done. We start by writing

$$\begin{aligned} L^* := \mathbb{P}[\text{(RP) is successful}] = \mathbb{P}[\exists \boldsymbol{y} \in [-\delta, \delta]^d : \boldsymbol{A}\boldsymbol{y} \in \mathcal{G}_U] = \mathbb{P}[\exists \boldsymbol{y} \in [-\delta, \delta]^d : \bar{\boldsymbol{B}}\boldsymbol{y} = \boldsymbol{z}] \\ = \mathbb{P}[\exists \boldsymbol{y} \in [-\bar{\delta}, \bar{\delta}]^d : \bar{\boldsymbol{B}}\boldsymbol{y} = \bar{\boldsymbol{z}}], \end{aligned} \tag{3.7}$$

where $\bar{\boldsymbol{B}}$ denotes a $d_e \times d$ Gaussian matrix and $\bar{\boldsymbol{z}} = \boldsymbol{z}/\mu$. Here, the second equality follows from Definition 1.3, and the third equality follows from Theorem 2.23 and the fact that $\mathcal{G}_U = \mathcal{G}_U^*$. Note that $\|\bar{\boldsymbol{z}}\| = 1$ since $\|\boldsymbol{z}\| = \|\boldsymbol{x}_\top^*\| = \mu$ (see Remark 2.24). We assign $\bar{\boldsymbol{z}}$ to a random vector with unit norm and keep $\bar{\boldsymbol{z}}$ fixed throughout the experiment[1]. For each $\bar{\delta} \in [0.02, 10]$, we generate 1000 Gaussian matrices $\bar{\boldsymbol{B}}$ and estimate $\mathbb{P}[\exists \boldsymbol{y} \in [-\bar{\delta}, \bar{\delta}]^d : \bar{\boldsymbol{B}}\boldsymbol{y} = \bar{\boldsymbol{z}}]$ in (3.7) as the proportion of instances for which the statement under the probability is true. Unlike R^*, L^* depends on individual values of d and d_e. We plot the estimates for L^* for the following values of $d - d_e$: 0, 1, 2, 3 and, in each plot, we repeat the experiment for $d_e = 2, 3, 4, 5, 6$. The plots are presented in Figure 3.2.

---

[1] Note that the results of the experiment are invariant of the choice of $\bar{\boldsymbol{z}}$ as long as its norm is fixed. Let $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ be two fixed vectors with unit norm. Consider two systems: $\boldsymbol{B}\boldsymbol{y} = \boldsymbol{z}_1$ and $\boldsymbol{B}\boldsymbol{y} = \boldsymbol{z}_2$. Note that $\boldsymbol{z}_2$ can be written as $\boldsymbol{Q}\boldsymbol{z}_1$ for some orthogonal $\boldsymbol{Q} \in \mathbb{R}^{d_e \times d_e}$. Then, the second system becomes $\boldsymbol{Q}^T\boldsymbol{B}\boldsymbol{y} = \boldsymbol{z}_1$ and this generates vectors $\boldsymbol{y}$ with the same distribution as the first system since $\boldsymbol{Q}^T\boldsymbol{B}$ is also Gaussian.

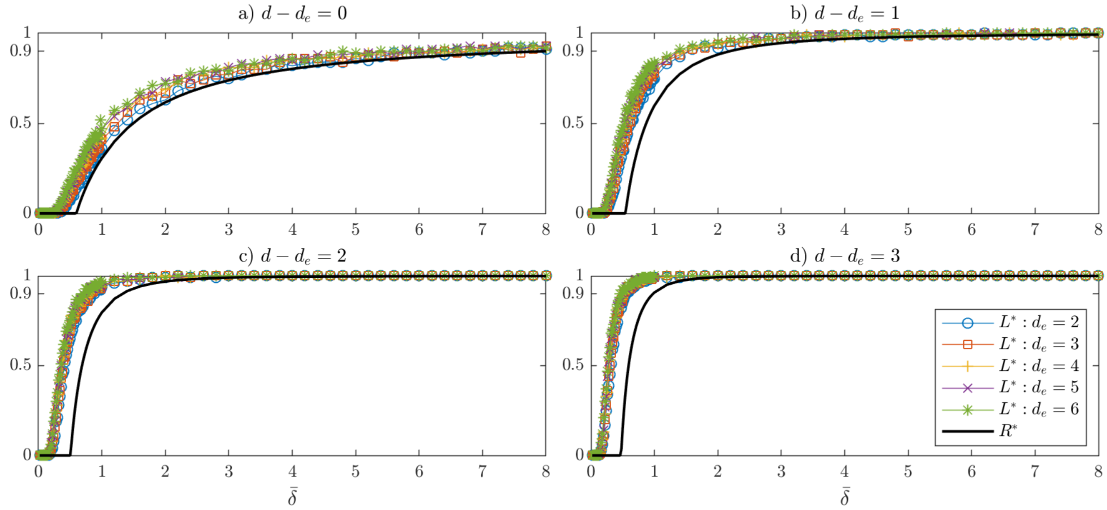


Figure 3.2: The four plots depict the function $R^*(\bar{\delta})$ and the estimates of $L^*(\bar{\delta})$; each plot corresponds to a particular value of $d - d_e \in \{0, 1, 2, 3\}$. Each plot contains estimates of $L^*(\bar{\delta})$ for $d_e = 2, 3, 4, 5, 6$.

Numerical findings. The plots in Figure 3.2 — confirming the conclusions of Corollary 3.4 — illustrate that the variation in success of (RP) is mainly determined by the value of $d - d_e$; the larger is the difference, the higher is the probability of success of (RP) for a given $\bar{\delta}$. These curves, being independent of μ, can be used to find suitable $\bar{\delta}$ for any problem for the corresponding values of $d - d_e$; the size of the $\mathcal{Y}$ box, δ, can then be set to $M\bar{\delta}$ if an upper bound M on μ is known.

Choosing d, δ and $\boldsymbol{p}$ in practice. When it comes to the numerical application of (RP) in practice, initialization of parameters d and δ might be problematic. From the theoretical discussions above we learned that the parameters d and δ must be defined based on d_e and μ, the values of which are typically unknown in practice, for example, for black-box functions. One could circumvent this issue by estimating d_e and μ rather than trying to calculate their exact values; note that all we need is an upper bound d on d_e. The parameter d_e or an upper bound may be known from prior studies or can be found with active subspace identification methods (see, e.g., [37]); these use gradients of f to estimate d_e.

Estimating μ can be a harder task. A rough estimate for μ can be obtained if the search in the original space is restricted to a certain domain; a trivial upper bound in this case is given by the maximum distance between $\boldsymbol{p}$ and the boundary of the domain. The search domain that is commonly imposed to practically solve unconstrained optimization problems is box constraints, such as $\mathcal{X} = [-1, 1]^D$ for which $\mu \leq \sqrt{D}$. In Section B.2, we test REGO assuming that $\sqrt{D}$ is the best bound known for μ. To compensate for unknown μ, one could also try increasing δ or d gradually to explore larger regions in $\mathbb{R}^D$.

As for $\boldsymbol{p}$, the theory indicates that choosing $\boldsymbol{p}$ closer to the set of global minimizers[2] increases the chances of successful recovery by (RP). If an approximate location of the set of global minimizers is known then one should choose $\boldsymbol{p}$ in that vicinity. Of course, in practice, the location of global minimizers is typically not known, therefore no preference is given to any particular $\boldsymbol{p}$. Numerically, one could solve multiple instances of (RP) in parallel each with a different $\boldsymbol{p}$ to increase coverage in $\mathbb{R}^D$ and, consequently, the overall probability of success.

3.3 Numerical methodology

This section contains the preliminary material for the numerical experiments in the present chapter and Chapter 6. We begin the section with the construction of functions with low-effective dimensionality, which form our test set. We then provide the descriptions of the three optimization solvers, namely, DIRECT, BARON and KNITRO, that we use to test (RP) and (RP$\mathcal{X}$).

Generating the test set. Our test set of functions with low effective dimensionality is derived from 19 global optimization problems (of dimensions 2–6) with known global minima [62, 49, 147], some of which are from the Dixon-Szego set [42]. The list of the problems is given in Table B.1, Section B.1.

Below we describe a technique adopted from Wang et al. [163] to generate high-dimensional functions with low effective dimensionality. Let $\bar{g}(\bar{\boldsymbol{x}})$ be any function from Table B.1; let d_e be its dimension and let the given domain be scaled to $[-1,1]^{d_e}$. We create a D-dimensional function $g(\boldsymbol{x})$ by adding $D-d_e$ fake dimensions to $\bar{g}(\bar{\boldsymbol{x}})$, $g(\boldsymbol{x}) = \bar{g}(\bar{\boldsymbol{x}}) + 0 \cdot x_{d_e+1} + 0 \cdot x_{d_e+2} + \cdots + 0 \cdot x_D$. We further rotate the function by applying a random orthogonal matrix $\boldsymbol{Q}$ to $\boldsymbol{x}$ to obtain a non-trivial constant subspace. The final form of the function we test is given as

$$f(\boldsymbol{x}) = g(\boldsymbol{Q}\boldsymbol{x}). \tag{3.8}$$

Note that the first d_e rows of $\boldsymbol{Q}$ now span the effective subspace $\mathcal{T}$ of $f(\boldsymbol{x})$. Furthermore, if the global minimum of $\bar{g}(\bar{\boldsymbol{x}})$ is attained[3] in $[-1,1]^{d_e}$, then $\min_{\bar{\boldsymbol{x}} \in \mathbb{R}^D} \bar{g}(\bar{\boldsymbol{x}}) \leq \sqrt{d_e}$, which when combined with (3.8) implies

$$\mu := \min_{\boldsymbol{x} \in \mathbb{R}^D} f(\boldsymbol{x}) \leq \sqrt{d_e}, \tag{3.9}$$

---

[2] Note that if $\boldsymbol{p}$ is closer to the set of global minimizers then μ is smaller.

[3] To the best of our knowledge, every function in Table B.1 attains its (unconstrained) global minimum inside its specified domain.

For each function in Table B.1, we generate three functions f as defined in (3.8) one for each $D = 10$, 100, 1000.

Optimization solvers. We test (RP) and (RP$\mathcal{X}$) with the below solvers.

DIRECT([58, 78, 52]) version 4.0 (DIviding RECTangles) is a deterministic[4] global optimization solver first introduced in [78] as an extension of Lipschitzian optimization. DIRECT does not require information about the gradient nor about the Lipschitz constant and, hence, can be used for black-box functions. DIRECT divides the search domain into rectangles and evaluates the function at the centre of each rectangle. Based on the previously sampled points, DIRECT carefully decides what rectangle to divide next balancing between local and global searches. Jones et al. [78] showed that DIRECT is guaranteed to converge to global minimum, but convergence may sometimes be slow.

BARON([139, 148]) version 17.10.10 (Branch-And-Reduce Optimization Navigator) is a branch-and-bound type global optimization solver for non-linear and mixed-integer non-linear programs. To provide lower and upper bounds for each branch, BARON utilizes algebraic structure of the objective function. It also includes a preprocessing step where it performs a multi-start local search to obtain a tight global upper bound. In comparison to other existing global solvers, BARON was demonstrated to be the most robust and fastest (see [119]). However, BARON accepts only a few (general) classes of functions[5] including polynomial, exponential, logarithmic, etc. and, unlike DIRECT, it is unable to optimize black-box functions.

KNITRO([28]) version 10.3.0 is a large-scale non-linear local optimization solver capable of handling problems with hundreds of thousands of variables. KNITRO allows to solve problems using one of the four algorithms: two interior point type methods (direct and conjugate gradient) and two active set type methods (active set and sequential quadratic programming). In contrast to BARON and DIRECT, which specialize on finding global minima, KNITRO focuses on finding local solutions. Nonetheless, KNITRO has multi-start capabilities, i.e., it solves a problem locally multiple times every time starting from a different point in the feasible domain. It is this feature that we make use of in the experiments.

Computational hardware and software. All the experiments conducted for this thesis were run in MATLAB on the 16 cores (2×8 Intel with hyper-threading) Linux machines with 256GB RAM and 3300 MHz speed.

---

[4] Here, we refer to the predictable behaviour of the solver given a fixed set of parameters.

[5] For instance, BARON cannot be applied to problems which include trigonometric functions.

Algorithm 1 Random Embeddings for Global Optimization (REGO) applied to (UP).

```
1: Initialise $\boldsymbol{p}$, $d$ and $\delta$ and define $\mathcal{Y} = [-\delta, \delta]^d$
2: Generate a $D \times d$ Gaussian matrix $\boldsymbol{A}$
3: Apply a global optimization solver (e.g. BARON, DIRECT, KNITRO) to (RP) until
   a termination criterion is satisfied, and define $\boldsymbol{y}_{min}$ to be the generated (approximate)
   solution of (RP).
4: Reconstruct $\boldsymbol{x}_{min} = \boldsymbol{A}\boldsymbol{y}_{min}$
```

3.4 Numerical experiments

We solve (UP) for each f in the test set both directly[6] (which we refer to here as '*no embedding*') and applying REGO (Random Embeddings for Global Optimization) outlined in Algorithm 1.

3.4.1 Setup

Experimental setup (REGO). We compare 'no embedding' and REGO using the three solvers described on page 51. Let g_i, s_j, n_j and D_k denote the ith function in the problem set (g_1 = Beale, etc., see Table B.1), jth solver (s_1 = DIRECT, s_2 = BARON, s_3 = KNITRO), the total number of problems in the problem set solvable by jth solver ($n_1 = 19$, $n_2 = 15$, $n_3 = 18$) and kth ambient dimension ($D_1 = 10$, $D_2 = 100$, $D_3 = 1000$), respectively. Let f_{ik} denote the D_k-dimensional function with low effective dimensionality constructed from g_i as described previously.

Within 'no embedding' framework, for each pair (s_j,D_k), we solve f_{ik} for $i = 1, 2, \ldots, n_j$ with solver s_j and record the proportion of the problems that attain convergence (see definition in Table 3.1).

For each f_{ik} ($1 \leq i \leq n_j$, $1 \leq k \leq 3$), we apply REGO 100 times every time with a different Gaussian matrix. Thus, in total, for each pair (s_j,D_k) we solve $n_j \times 100$ problems. We record the proportion of problems that attain convergence (see Table 3.1) out of these $n_j \times 100$ problems.

We also record the number of function evaluations (for DIRECT and KNITRO) and CPU time (for all the three solvers) spent before termination within the two frameworks. For each (s_j,D_k), function evaluations and time are averaged out over $n_j \times 100$ problems within REGO and over n_j problems within 'no embedding'.

We conduct the above experiment for REGO with $\boldsymbol{p} = \boldsymbol{0}$ and with the following pairs of parameters (d,δ): $(d_e, 8.0\times\sqrt{d_e})$, $(d_e+1, 2.2\times\sqrt{d_e})$, $(d_e+2, 1.3\times\sqrt{d_e})$ and $(d_e+3, 1.0\times\sqrt{d_e})$.

---

[6] To solve unconstrained (UP) numerically we need to impose bound constraints to limit the search. In no embedding setting, we solve (UP) over $\mathcal{X} = [-1, 1]^D$. The global minima of all the newly constructed functions are contained in $\mathcal{X}$.

Table 3.1: The table outlines the experimental setup for the three solvers. In the table, f is a function with low effective dimensionality d_e and the global minimum f^*, and ϵ is set to 10^{-3}.

	DIRECT	BARON	KNITRO
Measure of computational cost	function evaluations, CPU seconds	CPU seconds	function evaluations, CPU seconds
Budget per problem	$10000 \times d_e$ function evaluations	$200 \times d_e$ CPU seconds	$20 \times d_e$ starting points
Convergence criteria (see Remark 3.5)	$f^*_{Di} \leq f^* + \epsilon$	Convergence: $f^U_{Ba} \leq f^* + \epsilon$ Convergence$_{opt}$: $f^U_{Ba} \leq f^L_{Ba} + \epsilon$	$f^*_{Kn} \leq f^* + \epsilon$
Termination criteria	Either on budget or if $\boldsymbol{x}^*_{Di}$ satisfies the convergence criteria	Either on budget or if f^U_{Ba} and f^L_{Ba} satisfy the convergence$_{opt}$ criteria	On budget
Additional options	`options.testflag`=1 `options.maxits`=`Inf` `options.globalmin`=f^*	`npsol` = 9 `numloc` = 0 `BrVarStra` = 1 `BrPtStra` = 1	Default options. Derivatives are allowed. Use of multi-start through `ms_enable`=1.

Here, each δ was set to $M\bar{\delta}$, where $M = \sqrt{d_e}$ is an upper bound on μ (see (3.9)) and the value for $\bar{\delta}$ was chosen as the smallest $\bar{\delta}$ that gives at least 90% chance of success based on the curve of R^* in Figure 3.2.

Experimental setup (solvers). Due to the difference in algorithmic procedures of the three solvers we use, they allow different budget constraints and have different convergence and termination criteria; we present these in Table 3.1.

Remark 3.5. Every iteration of DIRECT stores f^*_{Di} — the minimum value of f found so far. BARON, at its every iteration, stores f^U_{Ba} and f^L_{Ba} — smallest upper bound and largest lower bound so far found for f. As for KNITRO, $f^*_{Kn} = \min\{f(\boldsymbol{A}\boldsymbol{y}^*_1), f(\boldsymbol{A}\boldsymbol{y}^*_2), \ldots, f(\boldsymbol{A}\boldsymbol{y}^*_l)\}$, where l is the number of starting points and where $\{\boldsymbol{y}^*_i\}_{1 \leq i \leq l}$ are the local solutions produced by the multi-start procedure.

Remark 3.6. The experiments are done not to compare solvers but to contrast 'no embedding' with REGO.

3.4.2 Numerical results

Success rates of (RP)**.** We record the proportion of instances for which (RP) is successful. Table 3.2 presents these percentages for each particular choice of d and D averaged over 19 problems in the test set. We observe that the percentages are very high and appear to be independent of D, supporting the conclusions of Corollary 3.4.

Table 3.2: The table shows average percentages of problems for which (RP) is successful.

d/D	10	100	1000
$d_e + 0$	97.2	97.8	97.3
$d_e + 1$	99.1	98.9	99.3
$d_e + 2$	99.5	99.6	99.8
$d_e + 3$	100	99.9	99.8

REGO vs. no embedding. The results of the experiments comparing REGO and 'no embedding' are presented in Figure 3.3, Figure 3.4 and Figure 3.5 for DIRECT, BARON and KNITRO, respectively. These figures compare average proportions of converged solutions and computational costs produced by REGO and 'no embedding' frameworks for $D = 10, 100, 1000$.

- DIRECT (Figure 3.3). For all the four initialisations of REGO, we observe that the average proportions of problems that attained convergence (see definition in Table 3.1) are invariant with respect to the ambient dimension. This frequency of convergence is higher within 'no embedding' for $D = 10, 100$, but exhibits a significant drop for $D = 1000$. The average function evaluation count is maintained within REGO, but doubles within 'no embedding' for a tenfold increase in D. Growth in CPU time takes place within both frameworks, being highest for 'no embedding'.

- BARON (Figure 3.4). In comparison with 'no embedding', the frequency of 'convergence$_{opt}$' is higher within REGO in most cases. We note that BARON's both 'convergence' and 'convergence$_{opt}$' exhibit invariance with respect to D within REGO. As for 'no embedding', we observe a decrease in the frequencies of both 'convergence' and 'convergence$_{opt}$'. In addition, we observe an increase in CPU time spent within 'no embedding', whilst the time is almost constant within REGO.

- KNITRO (Figure 3.5). We see that the proportion of solved problems is invariant with respect to the ambient dimension within REGO and, surprisingly, within 'no embedding' as well. However, the average number of function evaluations and time spent differ significantly between the two frameworks. With REGO, the average number of function evaluations remain at the same level for all D. Average time grows within both frameworks, but at a higher rate for 'no embedding'. The average time differs by a factor of 70 for $D = 1000$ in favour of REGO. We think that the growth in time within REGO is due to more costly function and derivative evaluations for larger D.

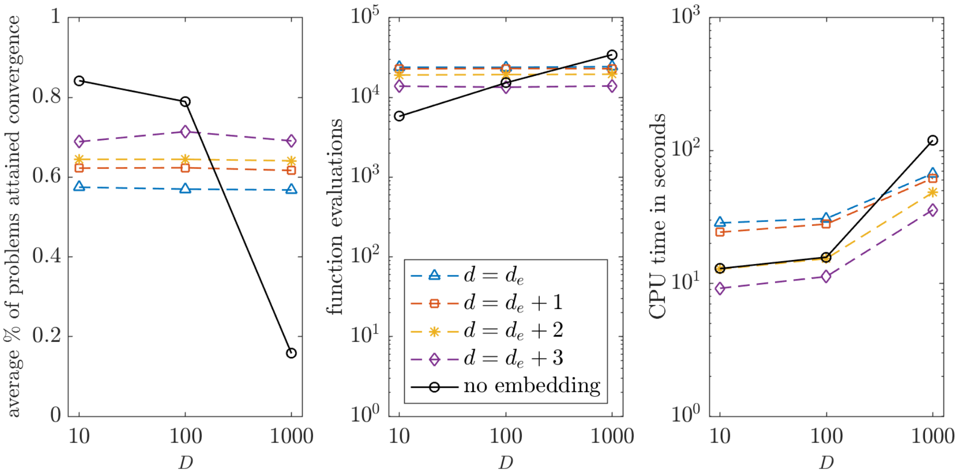


Figure 3.3: REGO versus 'no embedding' with DIRECT: comparison of frequency of convergence, average function evaluations (log scale) and average CPU time (in seconds, log scale).

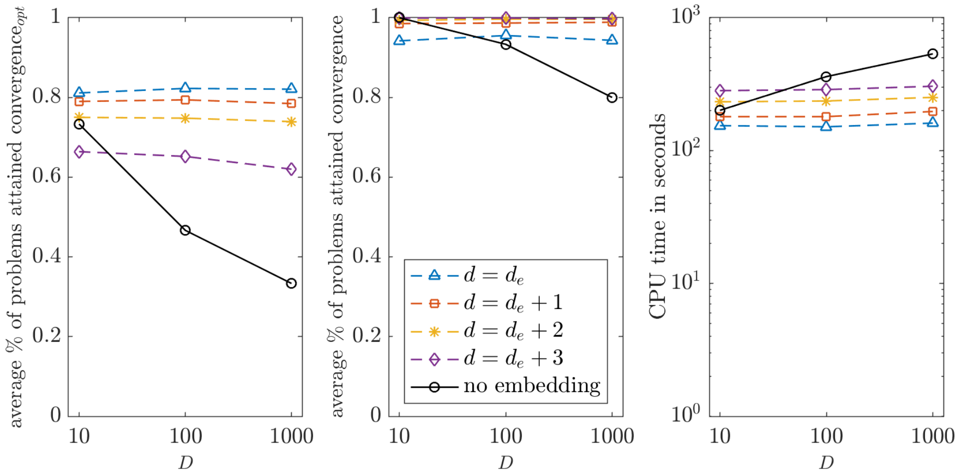


Figure 3.4: REGO versus 'no embedding' with BARON: comparison of frequency of convergence$_{opt}$/convergence and average CPU time (in seconds).

Summary of numerical results

1. (Effects of parameter choices) Our experiments clearly show that the choice of d and δ has a considerable effect on convergence and computational cost of REGO, and that good choices of (d, δ) are dependent on the underlying solver. For example, BARON achieved highest proportion of convergence$_{opt}$ within least amount of time for $(d, \delta) = (d_e, 8\sqrt{d_e})$, whereas DIRECT performed best for $(d_e + 3, \sqrt{d_e})$. KNITRO produced highest proportion of convergence and worst time for $(d_e + 3, \sqrt{d_e})$, and lowest proportion of convergence and best time for $(d_e, 8\sqrt{d_e})$.

2. (Scalability) Within REGO, the proportion of problems solved and/or number of

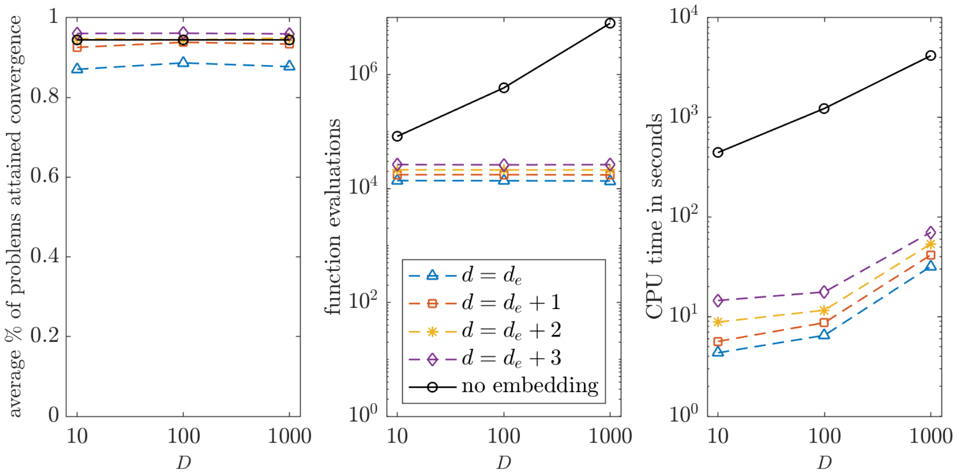


Figure 3.5: REGO versus 'no embedding' with KNITRO: comparison of frequency of convergence, log of average function evaluations and log of average CPU time (in seconds).

function evaluations are generally invariant with respect to the ambient dimension D. REGO displays good scalability for all three solvers.

3. (No embedding) Within 'no embedding', as D increases, the proportion of problems that attained convergence$_{opt}$/convergence decreased for BARON and DIRECT. Surprisingly, for the KNITRO's multi-start method, the proportion of solved problems is maintained, but the number of function evaluations and time increased dramatically.

Additional experiments. To see how robust REGO is to changes in the parameters, we conduct three more experiments presented and discussed in Section B.2. In the first experiment, assuming that μ is bounded by $\sqrt{D}$ (see page 50 for an explanation for this choice), we set δ to $\sqrt{D}\bar{\delta}$ for $\bar{\delta}$ chosen as in the main experiment. The second experiment tests REGO for four different values of d while keeping δ fixed and the third experiment tests REGO for three different values of δ keeping d fixed. In all three experiments, REGO performs well, particularly for BARON and KNITRO, solving most of the problems and exhibiting similar trends as in the main experiment.

3.5 Summary

In this chapter, we investigated the reduced problem (RP) associated to the unconstrained problem (UP) and for f with low effective dimensionality attempted to answer two questions:

1. How do parameters of (RP) affect the success of (RP)?
2. Is solving (RP) computationally as hard as solving the original (UP)?

In an attempt to answer these questions, we established some interesting results and made a few observations. In Section 3.1, we established a notable relationship between the success of (RP) and a chi-squared random variable. This result allowed us to obtain a quantifiable, interpretable lower bound on the success of (RP) which improves on the existing bounds in the literature. An important implication of this result was the fact that the success of (RP) has no explicit dependence on the ambient dimension D.

Furthermore, we proposed the REGO framework that is based on a single application of (RP). Numerical experiments with REGO supported our theoretical findings, demonstrating good scalability of (RP) with respect to a direct solution of (UP). We tested REGO for three state-of-the-art optimization solvers on a set of synthetically constructed functions with low effective dimensionality, which were derived from 19 classical global optimization test problems. For each solver, we performed the tests with different sets of parameters (d, δ) (values of which were determined using the derived theoretical lower bound in Corollary 3.4); the numerical results suggest that optimal combinations of parameters differ for different solvers and most importantly, in all cases/parameter combinations, the success rate of REGO is independent of (growing) ambient dimension.

The main limitation of REGO, however, is that it needs specification of parameters d and δ which are determined by the two problem constants that are typically unknown in practice, namely, the effective subspace dimension d_e and the distance to the closest minimizer μ. To specify suitable values for d and δ, it is sufficient to know upper bounds on d_e and μ, i.e., the knowledge of their exact values is not required. In this regard, one could estimate d_e using active subspace methods (see, e.g., [37]), which use derivatives, or numerically by increasing d incrementally until no significant decrease in the computed solution is observed. Obtaining an upper bound on μ is a more challenging issue. In this case, we suggest to set δ as large as one's computational budget allows, to cover more space in the ambient dimension.

Let us now move on to investigations of (RP$\mathcal{X}$) applied to constrained (P) for which we attempt to answer similar questions.

Chapter 4

Constrained optimization of functions with low effective dimensionality

In this chapter, we investigate the reduced problem (RP$\mathcal{X}$) associated to the constrained problem (P). The entire chapter is dedicated to derivations of lower bounds for the probability of success of (RP$\mathcal{X}$). The algorithm, its convergence and numerical performance pertaining to (RP$\mathcal{X}$) are presented in Chapter 6. Let us provide a roadmap of results that we present in this chapter.

Section 4.1 We begin the chapter with the derivation of the exact distribution of the minimizer $\boldsymbol{A}\boldsymbol{y}_2^* + \boldsymbol{p}$ in $\mathbb{R}^D$, where $\boldsymbol{y}_2^*$ is defined in (2.18), following a similar line of argument as in Section 2.4.1.

Section 4.2 Using the exact distribution of $\boldsymbol{A}\boldsymbol{y}_2^* + \boldsymbol{p}$ derived in Section 4.1, we establish a general lower bound for the probability of success of (RP$\mathcal{X}$). We then use this result to show that the probability of success of (RP$\mathcal{X}$) is positive and to derive quantifiable lower bounds for the special case of coordinate-aligned effective subspaces.

Section 4.3 Here, we introduce a notion of approximate success of (RP$\mathcal{X}$). This weaker notion of success with an additional Lipschitz continuity assumption on f allows us to obtain a positive lower bound that holds uniformly for all $\boldsymbol{p} \in \mathcal{X}$.

4.1 The distribution of the image of a reduced minimizer in $\mathbb{R}^D$

We first present a trivial lower bound for the probability of success of (RP$\mathcal{X}$) that is expressed in terms of a reduced minimizer $\boldsymbol{y}_2^*$ defined in (2.18) being feasible for the domain $\mathcal{X}$.

Lemma 4.1. *Suppose that Assumption LowED holds. Let* $\boldsymbol{x}^*$ *be a(ny) global minimizer of* (P)*,* $\boldsymbol{p} \in \mathcal{X}$*, a given vector, and* $\boldsymbol{A}$*, a* $D \times d$ *Gaussian matrix with* $d \geq d_e$*. Let* $\boldsymbol{y}_2^*$ *be defined in* (2.18)*. The reduced problem* (RP$\mathcal{X}$) *is successful in the sense of Definition 1.4 if* $\boldsymbol{A}\boldsymbol{y}_2^* + \boldsymbol{p} \in \mathcal{X}$*, that is*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \textit{ is successful}] \geq \mathbb{P}[-\mathbf{1} \leq \boldsymbol{A}\boldsymbol{y}_2^* + \boldsymbol{p} \leq \mathbf{1}]. \tag{4.1}$$

Proof. This is an immediate consequence of Definition 1.4 and (2.18), as the latter implies $\boldsymbol{A}\boldsymbol{y}_2^* + \boldsymbol{p} \in \mathcal{G}^*$ and so $f(\boldsymbol{A}\boldsymbol{y}_2^* + \boldsymbol{p}) = f^*$. ☐

To be able to quantify the lower bound in (4.1), let us rewrite $\boldsymbol{A}\boldsymbol{y}_2^* + \boldsymbol{p}$ in a different but equivalent form, expressed through only one random vector.

Let $\boldsymbol{Q} = (\boldsymbol{U} \;\; \boldsymbol{V})$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are defined in Assumption LowED. Since $\boldsymbol{Q}$ is orthogonal, for $\boldsymbol{A}\boldsymbol{y}_2^*$, we have

$$\boldsymbol{A}\boldsymbol{y}_2^* = \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{y}_2^* = \boldsymbol{Q}\begin{pmatrix}\boldsymbol{U}^T \\ \boldsymbol{V}^T\end{pmatrix}\boldsymbol{A}\boldsymbol{y}_2^*. \tag{4.2}$$

Using (2.19), we get $\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{y}_2^* = \boldsymbol{z}^*$. Letting

$$\boldsymbol{w} := \boldsymbol{V}^T\boldsymbol{A}\boldsymbol{y}_2^*, \tag{4.3}$$

we obtain

$$\boldsymbol{A}\boldsymbol{y}_2^* + \boldsymbol{p} = \boldsymbol{Q}\begin{pmatrix}\boldsymbol{z}^* \\ \boldsymbol{w}\end{pmatrix} + \boldsymbol{p} = \begin{pmatrix}\boldsymbol{U} & \boldsymbol{V}\end{pmatrix}\begin{pmatrix}\boldsymbol{z}^* \\ \boldsymbol{w}\end{pmatrix} + \boldsymbol{p} = \boldsymbol{U}\boldsymbol{z}^* + \boldsymbol{V}\boldsymbol{w} + \boldsymbol{p} = \boldsymbol{x}_\top^* + \boldsymbol{p}_\perp + \boldsymbol{V}\boldsymbol{w}, \tag{4.4}$$

where in the last equality we have used $\boldsymbol{U}\boldsymbol{z}^* = \boldsymbol{x}_\top^* - \boldsymbol{p}_\top$, see Theorem 2.23, and $\boldsymbol{p} = \boldsymbol{p}_\top + \boldsymbol{p}_\perp$, where $\boldsymbol{p}_\perp$ is the Euclidean projection of $\boldsymbol{p}$ on the constant subspace $\mathcal{T}^\perp$ of the objective function. Note that all the variables on the right-hand side of (4.4) are non-random except $\boldsymbol{w}$; thus, we can describe the distribution of $\boldsymbol{A}\boldsymbol{y}_2^* + \boldsymbol{p}$ solely through $\boldsymbol{w}$. In what follows, we derive the distribution of this random vector, which we will later utilize to obtain interpretable bounds for the success of (RP$\mathcal{X}$).

4.1.1 The distribution of the Euclidean norm

We derive the probability density function of the random vector $\boldsymbol{w}$ defined in (4.3) following a similar line of argument as in Section 2.4.1. We first derive the distribution of $\|\boldsymbol{w}\|$ and then show that $\boldsymbol{w}$ follows a spherical distribution, which then allows us to derive the exact distribution of $\boldsymbol{w}$.

Theorem 4.2 (Distribution of $\|\boldsymbol{w}\|_2^2$). *Suppose that Assumption LowED holds. Let* $\boldsymbol{x}^*$ *be a(ny) global minimizer of* (P), $\boldsymbol{p} \in \mathcal{X}$*, a given vector, and* $\boldsymbol{A}$*, a* $D \times d$ *Gaussian matrix with* $d \geq d_e$*. Assume that* $\boldsymbol{p}_\top \neq \boldsymbol{x}^*_\top$*. Let* $\boldsymbol{w}$ *be defined in* (4.3)*. Then,*

$$\left(\frac{1}{\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|_2^2} \cdot \frac{n}{m} \right) \|\boldsymbol{w}\|_2^2 \sim F(m,n),$$

where $m = D - d_e$*,* $n = d - d_e + 1$*, and where* $F(m,n)$ *denotes the* F*-distribution with degrees of freedom* m *and* n*.*

Proof. We write $\boldsymbol{w}$ as $\boldsymbol{C}\boldsymbol{y}_2^*$, where $\boldsymbol{C} = \boldsymbol{V}^T\boldsymbol{A}$. We first establish three facts: a) $\boldsymbol{B}$ and $\boldsymbol{C}$ are independent; b) $\boldsymbol{y}_2^*$ and $\boldsymbol{C}$ are independent; c) $\mathbb{P}[\boldsymbol{y}_2^* \neq \boldsymbol{0}] = 1$.

a) Since $\boldsymbol{V}$ is orthonormal, Theorem 2.2 implies that $\boldsymbol{C}$ is a Gaussian matrix. Moreover, the fact $\boldsymbol{U}^T\boldsymbol{V} = \boldsymbol{0}$ implies that $\boldsymbol{B}$ and $\boldsymbol{C}$ are independent, see Theorem 2.3.

b) Since $\boldsymbol{y}_2^*$ is measurable as a function of $\boldsymbol{B}$ (see proof of Theorem 2.29) and since $\boldsymbol{B}$ and $\boldsymbol{C}$ are independent, $\boldsymbol{y}_2^*$ and $\boldsymbol{C}$ must also be independent.

c) We have $\mathbb{P}[\boldsymbol{y}_2^* \neq \boldsymbol{0}] = 1 - \mathbb{P}[\boldsymbol{y}_2^* = \boldsymbol{0}] = 1 - \mathbb{P}[\|\boldsymbol{y}_2^*\|_2^2 = 0] = 1 - 0$, where the last equality is due to the fact that $\|\boldsymbol{y}_2^*\|_2^2$ follows the (appropriately scaled) inverse chi-squared distribution (Theorem 2.27), which is a continuous distribution.

Now, we apply Theorem 2.7 to obtain

$$\frac{\|\boldsymbol{w}\|_2^2}{\|\boldsymbol{y}_2^*\|_2^2} = \frac{(\boldsymbol{y}_2^*)^T\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{y}_2^*}{\|\boldsymbol{y}_2^*\|_2^2} \sim \chi^2_{D-d_e}, \tag{4.5}$$

which together with Theorem 2.27 yields

$$\frac{\|\boldsymbol{w}\|_2^2}{\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|_2^2} \sim \frac{\chi^2_{D-d_e}}{\chi^2_{d-d_e+1}}, \tag{4.6}$$

where $\chi^2_{D-d_e}$ and $\chi^2_{d-d_e+1}$ are independent[1]. Using the definition of the F-distribution (see Definition 2.12), we obtain the desired result. □

Using Theorem 4.2, it is straightforward to derive the p.d.f of $\|\boldsymbol{w}\|$.

Theorem 4.3 (The p.d.f. of $\|\boldsymbol{w}\|$). *Suppose that Assumption LowED holds. Let* $\boldsymbol{x}^*$ *be a(ny) global minimizer of* (P), $\boldsymbol{p} \in \mathcal{X}$*, a given vector, and* $\boldsymbol{A}$*, a* $D \times d$ *Gaussian matrix with* $d \geq d_e$*. Assume that* $\boldsymbol{p}_\top \neq \boldsymbol{x}^*_\top$*. The p.d.f.* $h(\hat{w})$ *of* $\|\boldsymbol{w}\|$*, with* $\boldsymbol{w}$ *defined in* (4.3)*, is given by*

$$h(\hat{w}) = \frac{2\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{\hat{w}^{m-1}}{\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|^m} \left(1 + \frac{\hat{w}^2}{\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|^2} \right)^{-(m+n)/2}, \tag{4.7}$$

where $m = D - d_e$*,* $n = d - d_e + 1$*, and where* Γ *is the usual gamma function.*

---

[1] Theorem 2.7 implies that $\boldsymbol{y}_2^* (= \|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\| / \chi^2_{d-d_e+1})$ and $\chi^2_{D-d_e}$ are independent; hence, $\chi^2_{D-d_e}$ and $\chi^2_{d-d_e+1}$ must also be independent.

Proof. Let $X \sim F(D - d_e, d - d_e + 1)$. Theorem 4.2 implies that

$$\|\boldsymbol{w}\| \overset{law}{=} K\sqrt{X}, \tag{4.8}$$

where

$$K = \|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\| \sqrt{\frac{D - d_e}{d - d_e + 1}}. \tag{4.9}$$

For the p.d.f. of $\|\boldsymbol{w}\|$, we have

$$h(\hat{w}) = \frac{d}{d\hat{w}}\,\mathbb{P}[\|\boldsymbol{w}\| < \hat{w}] = \frac{d}{d\hat{w}}\,\mathbb{P}[K\sqrt{X} < \hat{w}] = \frac{d}{d\hat{w}}\,\mathbb{P}[X < \hat{w}^2/K^2] = \frac{2\hat{w}}{K^2} f(\hat{w}^2/K^2), \tag{4.10}$$

where $f(x)$ denotes the p.d.f of an F-distributed random variable with degrees of freedom $m = D - d_e$ and $n = d - d_e + 1$. By substituting (2.7) in (4.10), we obtain the desired result. □

4.1.2 The probability density function

To derive the p.d.f. of $\boldsymbol{w}$ we will rely on the fact that $\boldsymbol{w}$ has a spherical distribution (see Definition 2.13), as we show next.

Theorem 4.4 ($\boldsymbol{w}$ has a spherical distribution). *Suppose that Assumption LowED holds. Let $\boldsymbol{x}^*$ be a(ny) global minimizer of* (P)*, $\boldsymbol{p} \in \mathcal{X}$, a given vector, and $\boldsymbol{A}$, a $D \times d$ Gaussian matrix with $d \geq d_e$. Assume that $\boldsymbol{p}_\top \neq \boldsymbol{x}^*_\top$. The random vector $\boldsymbol{w}$, defined in* (4.3)*, has a spherical distribution.*

Proof. Our proof is similar to the proof of Theorem 2.29. Let $\boldsymbol{S}$ be any $(D - d_e) \times (D - d_e)$ orthogonal matrix. To prove that $\boldsymbol{w}$ has a spherical distribution, we need to show that

$$\boldsymbol{w} \overset{law}{=} \boldsymbol{S}\boldsymbol{w}. \tag{4.11}$$

Using (2.19), we write $\boldsymbol{w} = \boldsymbol{C}\boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{z}^*$, where $\boldsymbol{C} = \boldsymbol{V}^T\boldsymbol{A}$ and $\boldsymbol{B} = \boldsymbol{U}^T\boldsymbol{A}$ are Gaussian matrices independent of one another by the point a) of the proof of Theorem 4.2. Let $h : \mathbb{R}^{Dd\times 1} \to \mathbb{R}^{(D-d_e)\times 1}$ be a vector-valued function defined as

$$h(\text{vec}[\boldsymbol{C}^T\ \boldsymbol{B}^T]) = \boldsymbol{C}\boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{z}^*, \tag{4.12}$$

where $\text{vec}[\boldsymbol{C}^T\ \boldsymbol{B}^T]$ denotes the vector of the concatenated columns of $(\boldsymbol{C}^T\ \boldsymbol{B}^T)$. We can express h as

$$h(\text{vec}[\boldsymbol{C}^T\ \boldsymbol{B}^T]) = \begin{pmatrix} \frac{p_1(\boldsymbol{C},\boldsymbol{B})}{q(\boldsymbol{B})} & \frac{p_2(\boldsymbol{C},\boldsymbol{B})}{q(\boldsymbol{B})} & \cdots & \frac{p_{D-d_e}(\boldsymbol{C},\boldsymbol{B})}{q(\boldsymbol{B})} \end{pmatrix}^T,$$

where $p_i(\boldsymbol{C}, \boldsymbol{B})$ for $1 \leq i \leq D - d_e$ are some polynomials in the entries of $\boldsymbol{C}$ and $\boldsymbol{B}$ and $q(\boldsymbol{B}) = \det(\boldsymbol{B}\boldsymbol{B}^T)$. Since q and p_i's are polynomials in Gaussian random variables, they are all measurable. Furthermore, since $\boldsymbol{B}$ is Gaussian, by Theorem 2.4, $\mathbb{P}[q = 0] = 0$; this implies that p_i/q is a measurable function for each $i = 1, 2, \ldots, D - d_e$ (see [165, Theorem 4.10]).

We have

$$\boldsymbol{w} = h(\mathrm{vec}[\boldsymbol{C}^T\ \boldsymbol{B}^T]) \text{ and } \boldsymbol{S}\boldsymbol{w} = h(\mathrm{vec}[(\boldsymbol{S}\boldsymbol{C})^T\ \boldsymbol{B}^T]). \tag{4.13}$$

From Theorem 2.2 it follows that $\boldsymbol{C} \overset{law}{=} \boldsymbol{S}\boldsymbol{C}$; hence $\mathrm{vec}[\boldsymbol{C}^T\ \boldsymbol{B}^T] \overset{law}{=} \mathrm{vec}[(\boldsymbol{S}\boldsymbol{C})^T\ \boldsymbol{B}^T]$. We can now apply Lemma A.7 to conclude that

$$\boldsymbol{w} = h(\mathrm{vec}[\boldsymbol{C}^T\ \boldsymbol{B}^T]) \overset{law}{=} h(\mathrm{vec}[(\boldsymbol{S}\boldsymbol{C})^T\ \boldsymbol{B}^T]) = \boldsymbol{S}\boldsymbol{w}. \qquad \square \tag{4.14}$$

We are now ready to derive the p.d.f. of $\boldsymbol{w}$.

Theorem 4.5 (The p.d.f. of $\boldsymbol{w}$). *Suppose that Assumption LowED holds. Let $\boldsymbol{x}^*$ be a(ny) global minimizer of* (P)*, $\boldsymbol{p} \in \mathcal{X}$, a given vector, and $\boldsymbol{A}$, a $D \times d$ Gaussian matrix with $d \geq d_e$. Assume that $\boldsymbol{p}_\top \neq \boldsymbol{x}^*_\top$. The random vector $\boldsymbol{w}$ defined in* (4.3) *follows a $(D - d_e)$-dimensional t-distribution with parameters $d - d_e + 1$ and $\frac{\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|^2}{d - d_e + 1}\boldsymbol{I}$, and with p.d.f. $g(\bar{\boldsymbol{w}})$ given by*

$$g(\bar{\boldsymbol{w}}) = \frac{1}{(\sqrt{\pi}\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|)^m} \left[\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{n}{2})}\right] \left(1 + \frac{\bar{\boldsymbol{w}}^T \bar{\boldsymbol{w}}}{\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|^2}\right)^{-(m+n)/2}, \tag{4.15}$$

where $m = D - d_e$ and $n = d - d_e + 1$.

Proof. We show that the p.d.f. of $\boldsymbol{w}$ is given by (4.15). The identification with the t-distribution follows from (2.5). Let us first show that $\mathbb{P}[\boldsymbol{w} = \boldsymbol{0}] = 0$. Let $X \sim F(D - d_e, d - d_e + 1)$. We have

$$\mathbb{P}[\boldsymbol{w} = \boldsymbol{0}] = \mathbb{P}[\|\boldsymbol{w}\|^2 = 0] = \mathbb{P}[X = 0], \tag{4.16}$$

where in the last equality we applied Theorem 4.2. Since the F-distributed X is a continuous random variable, the last probability in (4.16) is equal to zero.

Since $\mathbb{P}[\boldsymbol{w} = \boldsymbol{0}] = 0$ and $\boldsymbol{w}$ has a spherical distribution (Theorem 4.4), Theorem 2.15 implies that the p.d.f. $g(\bar{\boldsymbol{w}})$ of $\boldsymbol{w}$ satisfies

$$g(\bar{\boldsymbol{w}}) = \frac{\Gamma(m/2)}{2\pi^{m/2}} h(\|\bar{\boldsymbol{w}}\|)\|\bar{\boldsymbol{w}}\|^{1-m}, \tag{4.17}$$

where $h(\cdot)$ denotes the p.d.f. of $\|\boldsymbol{w}\|$. By substituting (4.7) into (4.17), we obtain the desired result. $\square$

4.2 Bounding the success of the reduced problem (RP$\mathcal{X}$)

We summarize the above analysis in the following corollary — the central result of this chapter.

Corollary 4.6. *Suppose that Assumption LowED holds. Let* $\boldsymbol{x}^*$ *be a(ny) global minimizer of* (P)*,* $\boldsymbol{p} \in \mathcal{X}$*, a given vector, and* $\boldsymbol{A}$*, a* $D \times d$ *Gaussian matrix with* $d \geq d_e$*. Assume that* $\boldsymbol{p}_\top \neq \boldsymbol{x}^*_\top$*. Then*

$$\mathbb{P}[(\text{RP}\mathcal{X}) \textit{ is successful}] \geq \mathbb{P}(-\boldsymbol{1} \leq \boldsymbol{x}^*_\top + \boldsymbol{p}_\perp + \boldsymbol{V}\boldsymbol{w} \leq \boldsymbol{1}), \tag{4.18}$$

where $\boldsymbol{w}$ *is a random vector that follows a* $(D - d_e)$*-dimensional* t*-distribution with parameters* $d - d_e + 1$ *and* $\frac{\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|^2}{d - d_e + 1}\boldsymbol{I}$*.*

Proof. The result follows from derivations (4.1)–(4.4) and Theorem 4.5. □

Remark 4.7. For the vector $\boldsymbol{w}$ to be well-defined, we require $d_e < D$ (see Assumption LowED). If $d_e = D$, then $d = D$; letting $\boldsymbol{Q} = \boldsymbol{I}$ and using $\boldsymbol{z}^* = \boldsymbol{x}^* - \boldsymbol{p}$ in (4.1)–(4.4), it is straightforward to see that $\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is successful}] = 1$.

Remark 4.8. Furthermore, for $\boldsymbol{w}$ to be a well-defined multivariate t-distributed random vector and for the proof of Corollary 4.6 to hold, we must ensure that $\boldsymbol{x}^*_\top \neq \boldsymbol{p}_\top$. Nonetheless, for the exceptional case $\boldsymbol{x}^*_\top = \boldsymbol{p}_\top$, $\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is successful}] = 1$ since $f(\boldsymbol{p}) = f(\boldsymbol{p}_\top + \boldsymbol{p}_\perp) = f(\boldsymbol{x}^*_\top + \boldsymbol{p}_\perp) = f^*$ which means that $\boldsymbol{p}$ is a global minimizer and that (RP$\mathcal{X}$) is successful for every embedding with $\boldsymbol{y}^* = \boldsymbol{0}$. Exactly the same exception was made in the analysis of (RP), see Remark 3.2.

Remark 4.9. Note that in our analysis, we assume (see Assumption LowED) that f is defined over whole $\mathbb{R}^D$ (and so $\boldsymbol{x}$ can be evaluated outside $\mathcal{X}$), however, in some practical applications f may be defined only over $\mathcal{X}$. If the domain of f is restricted to $\mathcal{X}$ our analysis still applies (with a minor tweak) . We adapt our analysis to this case by artificially defining f for certain points outside $\mathcal{X}$. We define f for $\boldsymbol{x} \notin \mathcal{X}$ that satisfy the following condition: there exists $\tilde{\boldsymbol{x}} \in \mathcal{X}$ and $\boldsymbol{c} \in \mathbb{R}^{D-d_e}$ such that $\boldsymbol{x} = \tilde{\boldsymbol{x}} + \boldsymbol{V}\boldsymbol{c}$. For such $\boldsymbol{x}$, we define $f(\boldsymbol{x}) := f(\tilde{\boldsymbol{x}})$. For example, in the analysis we use the value of $f(\boldsymbol{x}^*_\top)$, where $\boldsymbol{x}^*_\top$ may lie outside $\mathcal{X}$ and so $f(\boldsymbol{x}^*_\top)$ may not be defined. However, $\boldsymbol{x}^*_\top$ can be written as $\boldsymbol{x}^*_\top = \boldsymbol{x}^* + \boldsymbol{V}\boldsymbol{c}$, where $\boldsymbol{x}^* \in \mathcal{X}$ and $\boldsymbol{V}\boldsymbol{c} = -\boldsymbol{x}^*_\perp$ and so, by the proposed definition, $f(\boldsymbol{x}^*_\top) = f(\boldsymbol{x}^*) = f^*$.

4.2.1 Positive probability of success of the reduced problem (RP$\mathcal{X}$)

In this section, we prove that the probability of success of (RP$\mathcal{X}$) is positive. We note here that the probability can be equal to zero if the affine subspace of global minimizers $\mathcal{G}^*$ passes through the boundary of $\mathcal{X}$ and is oriented in a certain manner. To exclude these special cases from the analysis, below we define a new set $\bar{G}^*$ associated with $\mathcal{G}^*$ and with its simply connected set of feasible global minimizers G^*. In the analysis, the condition $\mathrm{Vol}(\bar{G}^*) > 0$ will ensure that the probability of (RP$\mathcal{X}$) is strictly greater than 0 for any $\boldsymbol{p} \in \mathcal{X}$.

We can express $G^* = \mathcal{G}^* \cap \mathcal{X} = \{\boldsymbol{x}_\top^* + \boldsymbol{V}\boldsymbol{g} : -\boldsymbol{1} \leq \boldsymbol{x}_\top^* + \boldsymbol{V}\boldsymbol{g} \leq \boldsymbol{1}, \boldsymbol{g} \in \mathbb{R}^{D-d_e}\}$, where $\mathcal{G}^*$ and G^* are defined in Definition 2.20. For each G^*, we define the corresponding set of "admissible" $(D - d_e)$-dimensional vectors as

$$\bar{G}^* := \{\boldsymbol{g} \in \mathbb{R}^{D-d_e} : -\boldsymbol{1} \leq \boldsymbol{x}_\top^* + \boldsymbol{V}\boldsymbol{g} \leq \boldsymbol{1}\}. \tag{4.19}$$

Note that the set G^* is $(D-d_e)$-dimensional if and only if the volume of the set $\bar{G}^*$ in $\mathbb{R}^{D-d_e}$, denoted by $\mathrm{Vol}(\bar{G}^*)$, is non-zero. In some particular cases, when the global minimizer $\boldsymbol{x}^*$ in Definition 2.20 is on the boundary of $\mathcal{X}$, the corresponding simply connected component G^* may be of dimension strictly lower than $(D - d_e)$ and, hence, $\mathrm{Vol}(\bar{G}^*) = 0$.

Definition 4.10. Let G^* and $\bar{G}^*$ be defined as in Definition 2.20 and (4.19), respectively. We say that G^* is non-degenerate if $\mathrm{Vol}(\bar{G}^*) > 0$.

We now show that, with this non-degeneracy assumption, the probability of success of (RP$\mathcal{X}$) is positive.

Theorem 4.11. *Suppose that Assumption LowED holds and suppose that there is a set G^* defined in Definition 2.20 that is non-degenerate according to Definition 4.10. Let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix with $d \geq d_e$. Then, for any $\boldsymbol{p} \in \mathcal{X}$,*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \textit{ is successful}] > 0. \tag{4.20}$$

Proof. We consider two cases, $\boldsymbol{p} \in G$ and $\boldsymbol{p} \in \mathcal{X} \setminus G$. Firstly, assume that $\boldsymbol{p} \in G$. Then, $\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] = 1$ since taking $\boldsymbol{y} = \boldsymbol{0}$ in (RP$\mathcal{X}$) yields $f(\boldsymbol{p}) = f^*$ (see Remark 4.8).

Assume now that $\boldsymbol{p} \in \mathcal{X} \setminus G$. Let $\boldsymbol{x}^*$ be a feasible global minimizer that lies in G^*. Using (4.18) with this particular $\boldsymbol{x}^*$ and noting that $\boldsymbol{p}_\perp = \boldsymbol{V}\boldsymbol{V}^T\boldsymbol{p}$ gives us

$$\begin{aligned}
\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] &\geq \mathbb{P}[-\boldsymbol{1} \leq \boldsymbol{x}_\top^* + \boldsymbol{V}(\boldsymbol{V}^T\boldsymbol{p} + \boldsymbol{w}) \leq \boldsymbol{1}] \\
&= \mathbb{P}[\boldsymbol{V}^T\boldsymbol{p} + \boldsymbol{w} \in \{\boldsymbol{g} \in \mathbb{R}^{D-d_e} : -\boldsymbol{1} \leq \boldsymbol{x}_\top^* + \boldsymbol{V}\boldsymbol{g} \leq \boldsymbol{1}\}] \\
&= \mathbb{P}[\boldsymbol{V}^T\boldsymbol{p} + \boldsymbol{w} \in \bar{G}^*] \\
&= \mathbb{P}[\boldsymbol{w} \in -\boldsymbol{V}^T\boldsymbol{p} + \bar{G}^*] \\
&= \int_{-\boldsymbol{V}^T\boldsymbol{p}+\bar{G}^*} g(\bar{\boldsymbol{w}})d\bar{\boldsymbol{w}},
\end{aligned} \tag{4.21}$$

where $\bar{G}^*$ is defined in (4.19) and where $g(\bar{\boldsymbol{w}})$ is the p.d.f. of $\boldsymbol{w}$ given in (4.15). The latter integral is positive since $g(\bar{\boldsymbol{w}}) > 0$ for any $\bar{\boldsymbol{w}} \in \mathbb{R}^{D-d_e}$ and since $\mathrm{Vol}(-\boldsymbol{V}^T\boldsymbol{p} + \bar{G}^*) = \mathrm{Vol}(\bar{G}^*) > 0$ (invariance of volumes under translations), where $\mathrm{Vol}(\bar{G}^*) > 0$ by the assumption of non-degeneracy of G^*. $\square$

Note that the proof of Theorem 4.11 illustrates that the success probability of (RP$\mathcal{X}$), though positive, depends on the choice of $\boldsymbol{p}$ [2]. Next, under additional problem assumptions, we derive lower bounds on the success probability of (RP$\mathcal{X}$) that are independent of $\boldsymbol{p}$ and/or quantifiable.

4.2.2 Quantifying the success probability of (RP$\mathcal{X}$) in the special case of coordinate-aligned effective subspace

Provided the effective subspace $\mathcal{T}$ is aligned with coordinate axes and without loss of generality, we can write the orthonormal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, whose columns span $\mathcal{T}$ and $\mathcal{T}^\perp$, as $\boldsymbol{U} = [\boldsymbol{I}_{d_e}\ \boldsymbol{0}]^T$ and $\boldsymbol{V} = [\boldsymbol{0}\ \boldsymbol{I}_{D-d_e}]^T$.

Theorem 4.12. *Let Assumption LowED hold with* $\boldsymbol{U} = [\boldsymbol{I}_{d_e}\ \boldsymbol{0}]^T$ *and* $\boldsymbol{V} = [\boldsymbol{0}\ \boldsymbol{I}_{D-d_e}]^T$. *Let* $\boldsymbol{x}^*$ *be a(ny) global minimizer of* (P), $\boldsymbol{p} \in \mathcal{X}$, *a given vector, and* $\boldsymbol{A}$, *a* $D \times d$ *Gaussian matrix with* $d \geq d_e$. *Assume that* $\boldsymbol{p}_\top \neq \boldsymbol{x}^*_\top$. *Then*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \textit{ is successful}] \geq \mathbb{P}[-\boldsymbol{1} - \boldsymbol{p}_{d_e+1:D} \leq \boldsymbol{w} \leq \boldsymbol{1} - \boldsymbol{p}_{d_e+1:D}], \tag{4.22}$$

where $\boldsymbol{w}$ *is a random vector that follows a* $(D-d_e)$*-dimensional t-distribution with parameters* $d - d_e + 1$ *and* $\frac{\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|^2}{d-d_e+1}\boldsymbol{I}$.

Proof. For $\boldsymbol{x}^* \in G^*$, we have

$$\boldsymbol{x}^*_\top = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}^* = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \boldsymbol{x}^* = \begin{pmatrix} \boldsymbol{x}^*_{1:d_e} \\ \boldsymbol{0} \end{pmatrix}. \tag{4.23}$$

Furthermore,

$$\boldsymbol{p}_\perp = \boldsymbol{V}\boldsymbol{V}^T\boldsymbol{p} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \boldsymbol{p} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{p}_{d_e+1:D} \end{pmatrix}.$$

Note that $\boldsymbol{x}^* \in [-1,1]^D$ implies that $\boldsymbol{x}^*_{1:d_e} \in [-1,1]^{d_e}$. Corollary 4.6 then yields

$$\begin{aligned} \mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] &\geq \mathbb{P}(-\boldsymbol{1} \leq \boldsymbol{x}^*_\top + \boldsymbol{p}_\perp + \boldsymbol{V}\boldsymbol{w} \leq \boldsymbol{1}) \\ &= \mathbb{P}\left[\begin{pmatrix} -\boldsymbol{1} \\ -\boldsymbol{1} \end{pmatrix} \leq \begin{pmatrix} \boldsymbol{x}^*_{1:d_e} \\ \boldsymbol{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{p}_{d_e+1:D} \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{I} \end{pmatrix} \boldsymbol{w} \leq \begin{pmatrix} \boldsymbol{1} \\ \boldsymbol{1} \end{pmatrix}\right] \\ (\text{since } \boldsymbol{x}^*_{1:d_e} \in [-1,1]^{d_e}) &= \mathbb{P}[-\boldsymbol{1} \leq \boldsymbol{p}_{d_e+1:D} + \boldsymbol{w} \leq \boldsymbol{1}], \end{aligned}$$

which immediately gives (4.22). $\square$

---

[2] When $\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\| \to 0$, the multivariate t-distribution in Corollary 4.6 becomes degenerate. Thus it is challenging to derive a lower bound on the integral (4.21) that is uniformly bounded away from zero with respect to $\boldsymbol{p}$.

Note that the right-hand side of (4.22) can be written as the integral of the p.d.f. of $\boldsymbol{w}$ over the hyperrectangular region $-\mathbf{1} - \boldsymbol{p}_{d_e+1:D} \leq \boldsymbol{w} \leq \mathbf{1} - \boldsymbol{p}_{d_e+1:D}$. Instead of directly computing this integral, we analyse its asymptotic behaviour for large D, assuming that d_e and d are fixed. We obtain the following main result, with its proof provided in Section 4.2.3.

Theorem 4.13. *Let Assumption LowED hold with* $\boldsymbol{U} = [\boldsymbol{I}_{d_e}\ \mathbf{0}]^T$ *and* $\boldsymbol{V} = [\mathbf{0}\ \boldsymbol{I}_{D-d_e}]^T$. *Let* d_e *and* d *be fixed with* $d \geq d_e$, *and let* $\boldsymbol{A}$ *be a* $D \times d$ *Gaussian matrix. For all* $\boldsymbol{p} \in \mathcal{X}$, *we have*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \textit{ is successful}] \geq \tau > 0, \tag{4.24}$$

where τ *satisfies*

$$\tau = \Theta\left(\frac{\log(D-d_e+1)^{\frac{d-1}{2}}}{2^{D-d_e}\cdot(D-d_e+1)^{d_e}}\right) \textit{ as } D \to \infty, \tag{4.25}$$

and the constants in $\Theta(\cdot)$ *are independent of* D.

Proof. See Section 4.2.3. ☐

The next result shows that, in the particular case when $\boldsymbol{p} = \mathbf{0}$, the center of the full-dimensional domain $\mathcal{X}$, the probability of success decreases at worst algebraically[3] with the ambient dimension D.

Theorem 4.14. *Let Assumption LowED hold with* $\boldsymbol{U} = [\boldsymbol{I}_{d_e}\ \mathbf{0}]^T$ *and* $\boldsymbol{V} = [\mathbf{0}\ \boldsymbol{I}_{D-d_e}]^T$. *Let* d_e *and* d *be fixed with* $d \geq d_e$, *and let* $\boldsymbol{A}$ *be a* $D \times d$ *Gaussian matrix. Let* $\boldsymbol{p} = \mathbf{0}$. *Then*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \textit{ is successful}] \geq \tau_{\mathbf{0}} > 0, \tag{4.26}$$

where

$$\tau_{\mathbf{0}} = \Theta\left(\frac{\log(D-d_e+1)^{\frac{d-1}{2}}}{(D-d_e+1)^{d_e}}\right) \textit{ as } D \to \infty, \tag{4.27}$$

and where the constants in $\Theta(\cdot)$ *are independent of* D.

Proof. See Section 4.2.3. ☐

Remark 4.15. Unlike Theorem 4.11, the above result does not require the assumption that there exists a non-degenerate G^*. As the effective subspace is aligned with the coordinate axes, the assumption is satisfied in this case since $\bar{G}^* = \{\boldsymbol{g} \in \mathbb{R}^{D-d_e} : -\mathbf{1} \leq \boldsymbol{x}^*_\top + \boldsymbol{V}\boldsymbol{g} \leq \mathbf{1}\} = [-1,1]^{D-d_e}$, as $\boldsymbol{V} = [\mathbf{0}\ \boldsymbol{I}_{D-d_e}]^T$ and the last $D - d_e$ components of the vector $\boldsymbol{x}^*_\top$ are zero; see the proof of Theorem 4.12.

---

[3] This simplification is due to the fact that when $\boldsymbol{p} = \mathbf{0}$, the factor 2^{D-d_e} in the denominator of (4.25) disappears.

Remark 4.16. The lower bounds on the probability of success of the reduced problem derived here and in the previous section are reasonably tight. We note for example that (4.18) and, consequently, (4.22) hold with equality if $d = d_e$ and $G = G^*$. Our numerical experiments in Chapter 6 also clearly illustrate that the success probability decreases with growing problem dimension D.

Remark 4.17. Our particular choice of asymptotic framework here is due to its practicality as well as to the ready-at-hand analysis of a similar integral to (4.29) in [166]. The scenario (d_e and d fixed, D large) is a familiar one in practice, where commonly, d_e is small compared to D, and d is limited by computational resources available to solve the reduced subproblem. Other asymptotic frameworks that could be considered in the future are $d_e = O(1)$, $d = O(\log(D))$ or $d_e = O(1)$, $d = \beta D$ where β is fixed. For more details on how to obtain asymptotic expansions similar to (4.25) and (4.27) for such choices of d_e and d, refer to [149, 166].

4.2.3 Proofs of results in Section 4.2.2

In this section, we prove Theorem 4.13 and Theorem 4.14. A crucial lemma is given first.

Lemma 4.18. *In the conditions of Theorem 4.13, we have*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] \geq I(\boldsymbol{p}, \Delta), \tag{4.28}$$

where $\Delta := \|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|$ *and*

$$I(\boldsymbol{p}, \Delta) := \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \left(\prod_{i=d_e+1}^{D} \frac{1}{\sqrt{2\pi}} \int_{s(-1-p_i)/\Delta}^{s(1-p_i)/\Delta} e^{-x^2/2} dx \right) s^{n-1} e^{-s^2/2} ds. \tag{4.29}$$

Proof. Theorem 4.12 implies that

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] \geq \mathbb{P}[-\boldsymbol{1} - \boldsymbol{p}_{d_e+1:D} \leq \boldsymbol{w} \leq \boldsymbol{1} - \boldsymbol{p}_{d_e+1:D}],$$

where $\boldsymbol{w}$ follows a $(D - d_e)$-dimensional t-distribution with parameters $n = d - d_e + 1$ and $\boldsymbol{\Sigma} = (\Delta^2/n)\boldsymbol{I}$. Let $s \sim \sqrt{\chi^2_n}$, $m = D - d_e$ and $Z_1, \dots, Z_m$ be i.i.d standard Gaussian random variables. By applying Theorem 2.11 (with $X = s^2$, $\nu = n$ and $\sigma = \Delta/\sqrt{n}$), we obtain

$$\boldsymbol{w} \overset{law}{=} \frac{\Delta}{s} \begin{pmatrix} Z_1 \\ \vdots \\ Z_m \end{pmatrix}, \tag{4.30}$$

Then, (4.30) yields

$$\begin{aligned} \mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] &\geq \mathbb{P}[-\boldsymbol{1} - \boldsymbol{p}_{d_e+1:D} \leq \boldsymbol{w} \leq \boldsymbol{1} - \boldsymbol{p}_{d_e+1:D}] \\ &= \mathbb{P}\left[\frac{s}{\Delta}(-1 - p_{d_e+1}) \leq Z_1 \leq \frac{s}{\Delta}(1 - p_{d_e+1}), \dots, \frac{s}{\Delta}(-1 - p_D) \leq Z_m \leq \frac{s}{\Delta}(1 - p_D)\right], \end{aligned} \tag{4.31}$$

Note that the last probability in (4.31) can be written as the integral of the joint probability distribution of $Z_1, \ldots, Z_m$ and s over the specified bounds. Leaving s as the last variable to be integrated, we can write

$$\begin{aligned}
\mathbb{P}\left[\frac{s}{\Delta}(-1-p_{d_e+1}) \leq Z_1 \leq \frac{s}{\Delta}(1-p_{d_e+1}), \ldots, \frac{s}{\Delta}(-1-p_D) \leq Z_m \leq \frac{s}{\Delta}(1-p_D)\right]& \\
= \int_0^\infty G(\boldsymbol{p}, \Delta, s) h(s) ds,& \quad (4.32)
\end{aligned}$$

where $G(\boldsymbol{p}, \Delta, s)$ is the integral of the joint p.d.f. of $Z_1, \ldots, Z_m$ integrated over the given bounds, i.e.,

$$\begin{aligned}
G(\boldsymbol{p}, \Delta, s) &= \int_{s(-1-p_{d_e+1})/\Delta}^{s(1-p_{d_e+1})/\Delta} \cdots \int_{s(-1-p_D)/\Delta}^{s(1-p_D)/\Delta} \frac{1}{(2\pi)^{m/2}} e^{-\frac{1}{2}(x_1^2+\cdots+x_m^2)} dx_1 \ldots dx_m \\
&= \prod_{i=d_e+1}^{D} \frac{1}{\sqrt{2\pi}} \int_{s(-1-p_i)/\Delta}^{s(1-p_i)/\Delta} e^{-x^2/2} dx,
\end{aligned} \tag{4.33}$$

and where $h(s)$ is the p.d.f. of s given by

$$h(s) = \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} s^{n-1} e^{-s^2/2}. \tag{4.34}$$

By combining (4.31) – (4.34), we obtain (4.28)–(4.29). □

It is easier to show Theorem 4.14 first, when $\boldsymbol{p} = \boldsymbol{0}$.

4.2.3.1 Proof of Theorem 4.14

The next result is a direct corollary of Lemma 4.18 when $\boldsymbol{p} = \boldsymbol{0}$, allowing us to replace $I(\boldsymbol{p}, \Delta)$ in (4.28) with a new integral $J_{m,n}(\Delta)$ that will be easier to manipulate.

Corollary 4.19. *In the conditions and notation of Lemma 4.18, let* $\boldsymbol{p} = \boldsymbol{0}$. *Then*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \textit{ is successful}] \geq J_{m,n}(\|\boldsymbol{x}_\top^*\|), \tag{4.35}$$

where

$$J_{m,n}(\Delta) := \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \left(\sqrt{\frac{2}{\pi}} \int_0^{s/\Delta} e^{-x^2/2} dx\right)^m s^{n-1} e^{-s^2/2} ds. \tag{4.36}$$

Proof. Let $\boldsymbol{p} = \boldsymbol{0}$. Then Lemma 4.18 implies that $\mathbb{P}[(\mathrm{RP}\mathcal{X})$ is successful$] \geq I(\boldsymbol{0}, \|\boldsymbol{x}_\top^*\|)$, where

$$\begin{aligned}
I(\boldsymbol{0}, \|\boldsymbol{x}_\top^*\|) &= \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \left(\prod_{i=d_e+1}^{D} \frac{1}{\sqrt{2\pi}} \int_{-s/\|\boldsymbol{x}_\top^*\|}^{s/\|\boldsymbol{x}_\top^*\|} e^{-x^2/2} dx\right) s^{n-1} e^{-s^2/2} ds \\
&= \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \left(\sqrt{\frac{2}{\pi}} \int_0^{s/\|\boldsymbol{x}_\top^*\|} e^{-x^2/2} dx\right)^m s^{n-1} e^{-s^2/2} ds \\
&= J_{m,n}(\|\boldsymbol{x}_\top^*\|).
\end{aligned} \tag{4.37}$$

□

The following theorem provides an asymptotic expansion of $J_{m,n}(\Delta)$ for large m, that has algebraic dependence on m.

Theorem 4.20. *Let $J_{m,n}(\Delta)$ be the integral defined in (4.36). Let n and Δ be fixed and let $r = (n + \Delta^2 - 2)/2$. If $r \neq 0$ then, for large m,*

$$J_{m,n}(\Delta) = \frac{C(n,\Delta)}{(m+1)^{\Delta^2}} \Big((\log(m+1))^r - \frac{r}{2} \log(\log(m+1)) \cdot (\log(m+1))^{r-1} + O((\log(m+1))^{r-1}) \Big), \quad (4.38)$$

where

$$C(n,\Delta) = \pi^{\frac{\Delta^2}{2}} \Delta^n \frac{\Gamma(\Delta^2)}{\Gamma(n/2)}.$$

If $r = 0$, then $J_{m,n}(\Delta) = J_{m,1}(1) = 1/(m+1)$.

Proof. See the proof in Section 4.2.3.3. $\square$

Proof of Theorem 4.14 Corollary 4.19 implies that

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is successful}] \geq I(\mathbf{0}, \|\boldsymbol{x}_\top^*\|) \geq J_{m,n}(\|\boldsymbol{x}_\top^*\|). \quad (4.39)$$

By definition of $\boldsymbol{x}_\top^*$, there exists $\boldsymbol{x}^* \in G$ such that $\boldsymbol{x}_\top^* = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}^*$ with $\boldsymbol{U} = [\boldsymbol{I}_{d_e};\ \mathbf{0}]$. Then $\boldsymbol{x}_\top^* = [\boldsymbol{x}_{1:d_e}^*; \mathbf{0}]$ which implies $\|\boldsymbol{x}_\top^*\| \leq \sqrt{d_e}$. By monotonic decrease of $J_{m,n}$ (see Lemma A.8), (4.39) yields

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is successful}] \geq J_{m,n}(\sqrt{d_e})$$

for all $\boldsymbol{x}^*, \boldsymbol{p} \in \mathcal{X}$ such that $\boldsymbol{x}_\top^* \neq \boldsymbol{p}_\top$. If $\boldsymbol{x}_\top^* = \boldsymbol{p}_\top$, then

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is successful}] = 1 \geq J_{m,n}(\sqrt{d_e}),$$

where the inequality follows from Lemma A.9. Thus, (4.24) is satisfied for $\tau_{\mathbf{0}} = J_{m,n}(\sqrt{d_e})$, and (4.27) follows from Theorem 4.20. $\square$

4.2.3.2 Proof of Theorem 4.13

Unlike the case $\boldsymbol{p} = \mathbf{0}$, we cannot rewrite directly the integral $I(\boldsymbol{p}, \Delta)$ in terms if the integral $J_{m,n}(\Delta)$ (i.e., Corollary 4.19 does not hold) for $\boldsymbol{p} \in \mathcal{X}$ arbitrary. However, we derive a lower bound on $I(\boldsymbol{p}, \Delta)$ in terms of the simpler integral $J_{m,n}(\Delta)$ that is valid for all $\boldsymbol{p} \in \mathcal{X}$.

Lemma 4.21. *For any $\boldsymbol{p} \in \mathcal{X}$ and for any $\Delta > 0$, we have*

$$I(\boldsymbol{p}, \Delta) \geq \frac{1}{2^m} J_{m,n}(\Delta/2).$$

Proof. Let us define the function

$$g(z,\Delta,s) = \frac{1}{\sqrt{2\pi}} \int_{s(-1-z)/\Delta}^{s(1-z)/\Delta} e^{-x^2/2} dx, \tag{4.40}$$

and note that

$$I(\boldsymbol{p},\Delta) = \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \left(\prod_{i=d_e+1}^{D} g(p_i,\Delta,s) \right) s^{n-1} e^{-s^2/2} ds. \tag{4.41}$$

Next we find the minimizers of $g(z,\Delta,s)$ over $z \in [-1,1]$. Introducing the notation $l(z,\Delta,s) := s(1-z)/\Delta$, and using Leibniz integral rule, we obtain

$$\begin{aligned} \frac{dg(z,\Delta,s)}{dz} &= e^{\frac{-l(z,\Delta,s)^2}{2}} \frac{d(l(z,\Delta,s))}{dz} - e^{\frac{-l(-z,\Delta,s)^2}{2}} \frac{d(-l(-z,\Delta,s))}{dz} \\ &= e^{\frac{-s^2(1-z)^2}{2\Delta^2}} \left(\frac{-s}{\Delta} \right) - e^{\frac{-s^2(-1-z)^2}{2\Delta^2}} \left(\frac{-s}{\Delta} \right) \\ &= \frac{s}{\Delta} e^{-\frac{s^2}{2\Delta^2}(1+z^2)} \left(e^{-\frac{s^2 z}{\Delta^2}} - e^{\frac{s^2 z}{\Delta^2}} \right). \end{aligned} \tag{4.42}$$

Hence, $dg(z,\Delta,s)/dz$ is equal to zero if and only if

$$e^{-\frac{s^2 z}{\Delta^2}} - e^{\frac{s^2 z}{\Delta^2}} = 0, \tag{4.43}$$

which occurs only at $z = 0$. The sign of $dg(z,\Delta,s)/dz$ changes from negative to positive at $z = 0$ implying that the function is concave and so $g(z,\Delta,s)$ attains its maximum at $z = 0$ and its minimum at the boundaries. Since $g(z,\Delta,s)$ is symmetric around $z = 0$, the minimum is attained at $z = \pm 1$. Thus, for all $z \in [-1,1]$,

$$g(z,\Delta,s) \geq g(-1,\Delta,s) = \frac{1}{\sqrt{2\pi}} \int_{-l(1,\Delta,s)}^{l(-1,\Delta,s)} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_0^{\frac{2s}{\Delta}} e^{-x^2/2} dx. \tag{4.44}$$

By combining (4.44) with (4.41), we obtain

$$\begin{aligned} I(\boldsymbol{p},\Delta) &\geq \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \left(\prod_{i=d_e+1}^{D} \frac{1}{\sqrt{2\pi}} \int_0^{\frac{2s}{\Delta}} e^{-x^2/2} dx \right) s^{n-1} e^{-s^2/2} ds \\ &= \frac{1}{2^m} \cdot \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \left(\sqrt{\frac{2}{\pi}} \int_0^{\frac{2s}{\Delta}} e^{-x^2/2} dx \right)^m s^{n-1} e^{-s^2/2} ds \\ &= \frac{1}{2^m} J_{m,n}(\Delta/2). \end{aligned} \tag{4.45}$$

□

Proof of Theorem 4.13. Lemma 4.18 and Lemma 4.21 provide

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is successful}] \geq I(\boldsymbol{p}, \Delta) \geq \frac{1}{2^m} J_{m,n}(\Delta/2). \tag{4.46}$$

Let us now show that $\Delta \leq 2\sqrt{d_e}$ for all $\boldsymbol{x}^*, \boldsymbol{p} \in [-1,1]^D$. Since $\boldsymbol{U} = [\boldsymbol{I}_{d_e}\ \boldsymbol{0}]^T$, for any global minimizer $\boldsymbol{x}^*$, we have $\boldsymbol{x}^*_\top = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}^* = [\boldsymbol{x}^*_{1:d_e};\ \boldsymbol{0}]$, and for any $\boldsymbol{p}$, we have $\boldsymbol{p}_\top = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{p} = [\boldsymbol{p}_{1:d_e};\ \boldsymbol{0}]$. Since $\boldsymbol{x}^*, \boldsymbol{p} \in [-1,1]^D$, there holds $\|\boldsymbol{x}^*_\top\| \leq \sqrt{d_e}$ and $\|\boldsymbol{p}_\top\| \leq \sqrt{d_e}$, and hence, $\Delta = \|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\| \leq \|\boldsymbol{x}^*_\top\| + \|\boldsymbol{p}_\top\| \leq 2\sqrt{d_e}$.

Using the fact that $J_{m,n}(\Delta)$ is a monotonically decreasing function (see Lemma A.8), (4.46) yields

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is successful}] \geq \frac{1}{2^m} J_{m,n}(\sqrt{d_e}) \tag{4.47}$$

for all $\boldsymbol{x}^*, \boldsymbol{p} \in \mathcal{X}$ such that $\boldsymbol{x}^*_\top \neq \boldsymbol{p}_\top$. If $\boldsymbol{x}^*_\top = \boldsymbol{p}_\top$, then

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is successful}] = 1 \geq \frac{1}{2^m} J_{m,n}(\sqrt{d_e}),$$

where the inequality follows from Lemma A.9. Thus, (4.24) is satisfied for $\tau = J_{m,n}(\sqrt{d_e})/2^m$, and (4.25) follows from Theorem 4.20. □

4.2.3.3 Proof of Theorem 4.20

We rewrite $J_{m,n}(\Delta)$ as follows

$$\begin{aligned}
J_{m,n}(\Delta) &= \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \left(\sqrt{\frac{2}{\pi}} \int_0^{s/\Delta} e^{-x^2/2} dx \right)^m s^{n-1} e^{-s^2/2} ds \\
&= \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \left(\frac{2}{\sqrt{\pi}} \int_0^{\frac{s}{\sqrt{2}\Delta}} e^{-x^2} dx \right)^m s^{n-1} e^{-s^2/2} ds \\
&= \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \operatorname{erf}^m \left(\frac{s}{\sqrt{2}\Delta} \right) s^{n-1} e^{-s^2/2} ds,
\end{aligned}$$

where $\operatorname{erf}(\cdot)$ denotes the usual error function. After making an appropriate transformation, the integral becomes

$$J_{m,n}(\Delta) = \frac{2\Delta^n}{\Gamma(\frac{n}{2})} \int_0^\infty \operatorname{erf}^m(s) s^{n-1} e^{-\Delta^2 s^2} ds \tag{4.48}$$

In [166, Section 2, Chapter 2], Wong derives an asymptotic expansion of a similar integral; our derivations are based on his technique.

As s varies from 0 to ∞, $\operatorname{erf}(s)$ increases monotonically from 0 to 1. So, for m large almost all the mass of the integrand is concentrated at ∞. We make the substitution $e^{-t} = \operatorname{erf}(s)$ to bring the integral to the form:

$$J_{m,n}(\Delta) = \frac{\sqrt{\pi}\Delta^n}{\Gamma(\frac{n}{2})} \int_0^\infty e^{Ks(t)^2} s(t)^{n-1} e^{-(m+1)t} dt, \tag{4.49}$$

where $K = 1 - \Delta^2$ and $s(t) = \text{erf}^{-1}(e^{-t})$. Due to monotonicity of erf, $s(t)$ is uniquely defined for every t. As erf varies from 0 to 1, t varies from ∞ to 0. So the mass of the transformed integrand is now concentrated around 0.

We will derive the asymptotic expansion for (4.49) in three steps:

1. First, we will derive the asymptotic expansion of $e^{Ks(t)^2}s(t)^{n-1}$.
2. Then, we will show that, for any $0 < c < 1$, the integral
$$\int_c^{\infty} e^{Ks(t)^2}s(t)^{n-1}e^{-(m+1)t}dt \tag{4.50}$$
is exponentially small.
3. Finally, we will derive the asymptotic expansion of
$$\int_0^{c} e^{Ks(t)^2}s(t)^{n-1}e^{-(m+1)t}dt. \tag{4.51}$$

Step 1

In order to derive the asymptotic expansion of $e^{Ks(t)^2}s(t)^{n-1}$, we first need to derive the asymptotic expansion of $s(t)$. The derivation of the asymptotic expansion of $s(t)$ can be found in Wong [166, Lemma 1, p. 67], who finds the first two terms of the expansion. We extend his result in the following lemma providing a fifth order expansion.

Lemma 4.22. *For small positive t, $s(t) = \text{erf}^{-1}(e^{-t})$ satisfies*

$$s(t)^2 = -\log(t) - \frac{1}{2}\log(-\log(t)) - \log(\sqrt{\pi}) + \frac{\log(-\log(t))}{4(-\log(t))} - \frac{\log(e/\sqrt{\pi})}{2(-\log(t))} + O\left(\frac{\log^2(-\log(t))}{(\log(t))^2}\right).$$

Proof. The asymptotic expansion of $\text{erf}(s)$ at infinity is given by

$$\text{erf}(s) \sim 1 - \frac{e^{-s^2}}{\sqrt{\pi}s}\left(1 - \frac{1}{2s^2} + \frac{3}{4s^4} - \cdots\right)$$

By writing $1 - e^{-t} = 1 - \text{erf}(s)$ and using Taylor's expansion for e^{-t} at 0, we obtain

$$t(1 + O(t)) = \frac{e^{-s^2}}{\sqrt{\pi}s}\left(1 - \frac{1}{2s^2} + \frac{3}{4s^4} - \cdots\right).$$

By taking logs on both sides and using the Taylor's expansion for $\log(1 + x)$, we have

$$\log(t) + O(t) = -s^2 - \log(\sqrt{\pi}) - \log(s) - \frac{1}{2s^2} + O\left(\frac{1}{s^4}\right). \tag{4.52}$$

The dominant terms are $\log(t)$ and s^2, hence

$$s^2 \sim -\log(t), \text{ as } t \to 0^+. \tag{4.53}$$

To obtain higher order approximations, we write

$$s(t)^2 = -\log(t) + \epsilon_1(t)$$

and substitute this into (4.52). We have

$$\begin{aligned}\log(t) + O(t) = \log(t) - \epsilon_1(t) - \log(\sqrt{\pi}) - \frac{1}{2}\log(-\log(t)) - \frac{1}{2}\log\left(1 + \frac{\epsilon_1(t)}{-\log(t)}\right) + \\ + O\left(\frac{1}{-\log(t) + \epsilon_1(t)}\right)\end{aligned} \tag{4.54}$$

Note that by (4.53), as $t \to 0^+$

$$\frac{\epsilon_1(t)}{-\log(t)} \to 0. \tag{4.55}$$

By using (4.55) in (4.54), we obtain

$$\epsilon_1(t) = -\frac{1}{2}\log(-\log(t)) - \log(\sqrt{\pi}) + o(1). \tag{4.56}$$

To obtain the following leading terms in the approximation we write

$$s^2(t) = -\log(t) - \frac{1}{2}\log(-\log(t)) - \log(\sqrt{\pi}) + \epsilon_2(t) \tag{4.57}$$

and repeat the above procedure. We substitute (4.57) into (4.52) and after a little manipulation obtain

$$\begin{aligned}O(t) = -\epsilon_2(t) - \frac{1}{2}\log\left(1 - \frac{1}{2}\frac{\log(-\log(t))}{-\log(t)} - \frac{\log(\sqrt{\pi})}{-\log(t)} + \frac{\epsilon_2(t)}{-\log(t)}\right) - \\ - \frac{1}{2}\cdot\frac{1}{-\log(t)}\cdot\frac{1}{1 - \frac{1}{2}\frac{\log(-\log(t))}{-\log(t)} - \frac{\log(\sqrt{\pi})}{-\log(t)} + \frac{\epsilon_2(t)}{-\log(t)}} + O((-\log(t))^2)\end{aligned} \tag{4.58}$$

Using the fact (by (4.56)) that $\epsilon_2(t) = o(1)$ and Taylor's expansions for $\log(1+x)$ and $1/(1-x)$, we obtain

$$\begin{aligned}O(t) = -\epsilon_2(t) - \frac{1}{2}\left(-\frac{1}{2}\frac{\log(-\log(t))}{-\log(t)} + O\left(\frac{1}{-\log(t)}\right)\right) - \\ - \frac{1}{2}\cdot\frac{1}{-\log(t)}\left(1 + O\left(\frac{\log(-\log(t))}{-\log(t)}\right)\right),\end{aligned}$$

which yields

$$\epsilon_2(t) = \frac{\log(-\log(t))}{4(-\log(t))} + O\left(\frac{1}{-\log(t)}\right). \tag{4.59}$$

To obtain the following leading terms in the expansion of $\epsilon_2(t)$, we use (4.59) in (4.58) leaving the first term $(-\epsilon_2(t))$ as is:

$$\begin{aligned}O(t) = -\epsilon_2(t) - \frac{1}{2}\log\left(1 - \frac{1}{2}\frac{\log(-\log(t))}{-\log(t)} - \frac{\log(\sqrt{\pi})}{-\log(t)} + O\left(\frac{\log(-\log(t))}{(-\log(t))^2}\right)\right) - \\ - \frac{1}{2}\cdot\frac{1}{-\log(t)}\cdot\frac{1}{1 - \frac{1}{2}\frac{\log(-\log(t))}{-\log(t)} - \frac{\log(\sqrt{\pi})}{-\log(t)} + O\left(\frac{\log(-\log(t))}{(-\log(t))^2}\right)} + O((-\log(t))^2)\end{aligned}$$

Now, using Taylor's expansions for $\log(1+x)$ and $1/(1-x)$, we obtain

$$O(t) = -\epsilon_2(t) - \frac{1}{2}\left(-\frac{1}{2}\frac{\log(-\log(t))}{-\log(t)} - \frac{\log(\sqrt{\pi})}{-\log(t)} + O\left(\frac{\log^2(-\log(t))}{(-\log(t))^2}\right)\right) - \\ -\frac{1}{2}\cdot\frac{1}{-\log(t)}\left(1 + O\left(\frac{\log(-\log(t))}{-\log(t)}\right)\right),$$

Hence,

$$\epsilon_2(t) = \frac{\log(-\log(t))}{4(-\log(t))} - \frac{\log(e/\sqrt{\pi})}{2(-\log(t))} + O\left(\frac{\log^2(-\log(t))}{(-\log(t))^2}\right).$$

□

Using Lemma 4.22, it is straightforward to derive the asymptotic expansion of $e^{Ks(t)^2}s(t)^{n-1}$.

Corollary 4.23. *Let $l(t) = -\log(t)$. Then, as $t \to 0^+$,*

$$e^{Ks(t)^2}s(t)^{n-1} = e^{Kl(t)}\pi^{-K/2}l(t)^{\frac{n-1-K}{2}}\left(1 - \left(\frac{n-1-K}{4}\right)\frac{\log(l(t))}{l(t)} - \frac{\log(e^{K/2}\pi^{\frac{n-1-K}{4}})}{l(t)} \\ + O\left(\frac{\log^2(l(t))}{l(t)^2}\right)\right) \tag{4.60}$$

Proof. From Theorem 4.22 it follows that

$$e^{Ks(t)^2} = e^{Kl(t)}l(t)^{-K/2}\pi^{-K/2}\exp\left(\frac{K\log(l(t))}{4l(t)} - \frac{K\log(e/\sqrt{\pi})}{2l(t)} + O\left(\frac{\log^2(l(t))}{l(t)^2}\right)\right).$$

By using Taylor's expansion for exp we obtain

$$e^{Ks(t)^2} = e^{Kl(t)}l(t)^{-K/2}\pi^{-K/2}\left(1 + \frac{K\log(l(t))}{4l(t)} - \frac{K\log(e/\sqrt{\pi})}{2l(t)} + O\left(\frac{\log^2(l(t))}{l(t)^2}\right)\right). \tag{4.61}$$

Similarly, using Theorem 4.22 and binomial expansion, for $s(t)^{n-1}$, we have

$$(s(t)^2)^{\frac{n-1}{2}} = l(t)^{\frac{n-1}{2}}\left(1 - \frac{(n-1)\log(l(t))}{4l(t)} - \frac{(n-1)\log(\sqrt{\pi})}{2l(t)} + O\left(\frac{\log^2(l(t))}{l(t)^2}\right)\right) \tag{4.62}$$

By multiplying the leading terms in (4.61) and (4.62), we obtain the desired result. □

Step 2

Let us now show that, for any $0 < c < 1$, the integral given in (4.50) is exponentially small. More precisely, we will show that, for large m,

$$\int_c^\infty e^{Ks(t)^2}s(t)^{n-1}e^{-(m+1)t}dt = O\left(\frac{e^{-c(m+n)}}{m+n}\right). \tag{4.63}$$

Let $\operatorname{erf}(s) = e^{-t}$. First, we establish that

$$\text{there exists a positive constant } A \text{ such that } s(t) = \operatorname{erf}^{-1}(e^{-t}) \leq Ae^{-t} \text{ for all } t \in [c,\infty). \tag{4.64}$$

Note that (4.64) holds if there exists an $A > 0$ such that $\mathrm{erf}^{-1}(x) \leq Ax$ for all $x \in [0, e^{-c}]$. To prove this, we apply the Mean Value Theorem to erf^{-1} over $[0, x]$; thus there exists $y \in (0, x)$ such that

$$\frac{\mathrm{erf}^{-1}(x) - \mathrm{erf}^{-1}(0)}{x - 0} = (\mathrm{erf}^{-1})'(y) \tag{4.65}$$

Using the following formula for the derivative of the inverse of the error function [3, eq (2.4), p. 192],

$$(\mathrm{erf}^{-1}(x))' = \frac{\sqrt{\pi}}{2} e^{(\mathrm{erf}^{-1}(x))^2},$$

from (4.65), we obtain

$$\frac{\mathrm{erf}^{-1}(x)}{x} = \frac{\sqrt{\pi}}{2} e^{(\mathrm{erf}^{-1}(y))^2}. \tag{4.66}$$

Since erf^{-1} is an increasing function and $y < x \leq e^{-c}$, (4.66) gives

$$\mathrm{erf}^{-1}(x) \leq \frac{\sqrt{\pi}}{2} e^{(\mathrm{erf}^{-1}(e^{-c}))^2} x,$$

which proves (4.64).

Now, since $s(t)$ is a monotonically decreasing function with $s(\infty) = 0$, we have[4]

$$e^{Ks(t)^2} \leq \max\{1, e^{Ks(c)^2}\} \text{ for } t \geq c. \tag{4.67}$$

Using (4.64) and (4.67), we finally obtain

$$\begin{aligned}\int_c^\infty e^{Ks(t)^2} s(t)^{n-1} e^{-(m+1)t} dt &\leq A^{n-1} \max\{1, e^{Ks(c)^2}\} \int_c^\infty e^{-(m+n)t} dt \\ &= A^{n-1} \max\{1, e^{Ks(c)^2}\} \frac{e^{-c(m+n)}}{m+n}.\end{aligned}$$

Step 3

We now provide the asymptotic expansion of the integral given in (4.51). Using the expansion of $e^{Ks(t)^2} s(t)^{n-1}$ provided in Corollary 4.23, we obtain

$$\begin{aligned}\int_0^c e^{Ks(t)^2} s(t)^{n-1} e^{-(m+1)t} dt =&\pi^{-K/2} L\left(1-K, \frac{n-1-K}{2}, m+1\right) \\ &- \pi^{-K/2}\left(\frac{n-1-K}{4}\right) G\left(1-K, \frac{n-3-K}{2}, m+1\right) \\ &- \pi^{-K/2} \log(e^{K/2} \pi^{\frac{n-1-K}{4}}) L\left(1-K, \frac{n-3-K}{2}, m+1\right) + \cdots,\end{aligned} \tag{4.68}$$

---

[4] Over $t \in [c, \infty)$, for $K \geq 0$, $e^{Ks(t)^2} \leq e^{Ks(c)^2}$ and, for $K < 0$, $e^{Ks(t)^2} \leq 1$.

where $L(\cdot,\cdot,\cdot)$ and $G(\cdot,\cdot,\cdot)$ are defined as follows:

$$L(\lambda,\mu,z) = \int_0^c t^{\lambda-1}(-\log(t))^{\mu} e^{-zt}dt, \tag{4.69}$$

$$G(\lambda,\mu,z) = \int_0^c t^{\lambda-1}(-\log(t))^{\mu}\log(-\log(t)) e^{-zt}dt \tag{4.70}$$

for $0 < c < 1$. We derive the asymptotic expansions of $L(\lambda,\mu,z)$ and $G(\lambda,\mu,z)$ for large z in Theorems A.11 and A.10, respectively. By substituting the asymptotic expansions of $L(\lambda,\mu,z)$ and $G(\lambda,\mu,z)$ in (4.68), we obtain

$$\begin{aligned}\int_0^c e^{Ks(t)^2} s(t)^{n-1} e^{-(m+1)t}dt = \\ \frac{\pi^{-K/2}\Gamma(1-K)}{(m+1)^{1-K}}\Big((\log(m+1))^r - \frac{r}{2}\log(\log(m+1))\cdot(\log(m+1))^{r-1} \\ + O((\log(m+1))^{r-1})\Big),\end{aligned} \tag{4.71}$$

where $r = (n-K-1)/2$.

Summary

Let us now conclude the proof of Theorem 4.20. From (4.48), (4.49) and (4.63), it follows that

$$\begin{aligned}J_{m,n}(\Delta) &= \frac{2\Delta^n}{\Gamma(\frac{n}{2})}\int_0^\infty \operatorname{erf}^m(s) s^{n-1} e^{-\Delta^2 s^2} ds \\ &= \frac{\sqrt{\pi}\Delta^n}{\Gamma(\frac{n}{2})}\int_0^\infty e^{Ks(t)^2} s(t)^{n-1} e^{-(m+1)t}dt \\ &= \frac{\sqrt{\pi}\Delta^n}{\Gamma(\frac{n}{2})}\int_0^c e^{Ks(t)^2} s(t)^{n-1} e^{-(m+1)t}dt + O\left(\frac{e^{-c(m+n)}}{m+n}\right).\end{aligned} \tag{4.72}$$

By using (4.71) in (4.72) and substituting $K = 1-\Delta^2$, we obtain (4.38). Note that if $r = 0$ then $n = 1$ and $\Delta = 1$ and so $K = 0$. In this case, $e^{Ks(t)^2}s(t)^{n-1} = 1$ and direct integration yields $J_{m,1}(1) = 1/(m+1)$.

4.3 Uniformly positive lower bound on the ϵ-success probability of (RP$\mathcal{X}$) in the general case

As mentioned in the last paragraph of Section 4.2.1, it is difficult to derive a uniformly positive lower bound on the probability of success of (RP$\mathcal{X}$) that does not depend on $\boldsymbol{p}$. However, assuming Lipschitz continuity of the objective function, we are able to achieve such a guarantee for (RP$\mathcal{X}$) to be *approximately* successful, a weaker notion that is defined as follows.

Definition 4.24. For a(ny) $\epsilon > 0$, we say that (RP$\mathcal{X}$) is *ϵ-successful* if there exists $\boldsymbol{y}^* \in \mathbb{R}^d$ such that $f(\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}) \leq f^* + \epsilon$ and $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} \in \mathcal{X}$.

Let

$$G_\epsilon := \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) \leq f^* + \epsilon\} \tag{4.73}$$

be the set of feasible ϵ-minimizers. The reduced problem (RP$\mathcal{X}$) is thus ϵ-successful if it contains a feasible ϵ-minimizer.

Assumption LipC. The objective function $f : \mathbb{R}^D \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant L, that is, $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^D$.

The next theorem shows that the probability that (RP$\mathcal{X}$) is ϵ-successful is uniformly bounded away from zero for all $\boldsymbol{p} \in \mathcal{X}$.

Theorem 4.25 (Uniform bound). *Suppose that Assumptions LowED and LipC hold. Suppose also that there is a set G^* defined in Definition 2.20 that is non-degenerate according to Definition 4.10. Let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix with $d \geq d_e$ and $\epsilon > 0$, an accuracy tolerance. Then there exists a constant $\tau_\epsilon > 0$ such that, for all $\boldsymbol{p} \in \mathcal{X}$,*

$$\mathbb{P}[(\text{RP}\mathcal{X}) \textit{ is } \epsilon\textit{-successful}] \geq \tau_\epsilon. \tag{4.74}$$

Proof. Let $\boldsymbol{x}^*$ be a feasible global minimizer that lies in G^*. Let $N_\eta(G^*) := \{\boldsymbol{x} \in \mathcal{X} : \|\boldsymbol{x}^*_\top - \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}\|_2 \leq \eta\}$ be a neighbourhood of G^* in $\mathcal{X}$, for some $\eta > 0$, where as usual, $\boldsymbol{x}^*_\top = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}^*$ is the Euclidean projection of $\boldsymbol{x}^*$ on the effective subspace.

Firstly, assume that $\boldsymbol{p} \in N_{\epsilon/L}(G^*)$. Then, $\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\| \leq \epsilon/L$, and by Assumption LipC,

$$|f(\boldsymbol{p}) - f^*| = |f(\boldsymbol{p}_\top) - f(\boldsymbol{x}^*_\top)| \leq L\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\| \leq \epsilon.$$

Thus $\boldsymbol{p} \in G_\epsilon$ and, hence, $\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] = 1$. Otherwise, $\boldsymbol{p} \in \mathcal{X} \backslash N_{\epsilon/L}(G^*)$. Using the proof of Theorem 4.11, we have

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] \geq \mathbb{P}[(\text{RP}\mathcal{X}) \text{ is successful}] \geq \int_{-\boldsymbol{V}^T\boldsymbol{p}+\bar{G}^*} g(\bar{\boldsymbol{w}})d\bar{\boldsymbol{w}}, \tag{4.75}$$

where $\bar{G}^*$ is defined in (4.19), $g(\bar{\boldsymbol{w}})$ is the p.d.f. of $\boldsymbol{w}$ given by (4.15), and where the first inequality is due to the fact that (RP$\mathcal{X}$) being successful implies that (RP$\mathcal{X}$) is ϵ-successful (by letting $\epsilon := 0$ in Definition 4.24). To prove (4.74), it is thus sufficient to lower bound $g(\bar{\boldsymbol{w}})$ by a positive constant, independent of $\boldsymbol{p}$. Since, $\boldsymbol{p} \notin N_{\epsilon/L}(G^*)$, we have

$$\frac{\epsilon}{L} < \|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|_2 = \|\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{x}^* - \boldsymbol{p})\|_2 \leq \|\boldsymbol{U}\boldsymbol{U}^T\|_2 \cdot \|\boldsymbol{x}^* - \boldsymbol{p}\|_2 \leq 2\sqrt{D}, \tag{4.76}$$

where the last inequality follows from $\|\boldsymbol{U}\boldsymbol{U}^T\|_2 = 1$, since $\boldsymbol{U}$ has orthonormal columns, and from $-\mathbf{2} \le \boldsymbol{x}^* - \boldsymbol{p} \le \mathbf{2}$ since $\boldsymbol{x}^*, \boldsymbol{p} \in [-1, 1]^D$. Furthermore, note that, for any $\bar{\boldsymbol{w}} \in -\boldsymbol{V}^T\boldsymbol{p} + \bar{G}^*$, we have

$$-\mathbf{1} - \boldsymbol{x}^*_\top - \boldsymbol{p}_\perp \le \boldsymbol{V}\bar{\boldsymbol{w}} \le \mathbf{1} - \boldsymbol{x}^*_\top - \boldsymbol{p}_\perp,$$

and, hence,

$$\begin{aligned}
\|\boldsymbol{V}\bar{\boldsymbol{w}}\|_\infty &\le \max(\|-\mathbf{1} - \boldsymbol{x}^*_\top - \boldsymbol{p}_\perp\|_\infty, \|\mathbf{1} - \boldsymbol{x}^*_\top - \boldsymbol{p}_\perp\|_\infty) \\
&\le \|\mathbf{1}\|_\infty + \|\boldsymbol{x}^*_\top\|_\infty + \|\boldsymbol{p}_\perp\|_\infty \\
&\le 1 + \|\boldsymbol{x}^*_\top\|_2 + \|\boldsymbol{p}_\perp\|_2 \\
&= 1 + \|\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}^*\|_2 + \|\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{p}\|_2 \\
&\le 1 + \|\boldsymbol{U}\boldsymbol{U}^T\|_2 \cdot \|\boldsymbol{x}^*\|_2 + \|\boldsymbol{V}\boldsymbol{V}^T\|_2 \cdot \|\boldsymbol{p}\|_2 \\
&\le 1 + 2\sqrt{D},
\end{aligned}$$

where the last inequality follows from $\|\boldsymbol{U}\boldsymbol{U}^T\|_2 = 1$ and $\|\boldsymbol{V}\boldsymbol{V}^T\|_2 = 1$ (as $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal) and from $\boldsymbol{x}^*, \boldsymbol{p} \in [-1, 1]^D$. Thus,

$$\|\bar{\boldsymbol{w}}\|_2 = \|\boldsymbol{V}\bar{\boldsymbol{w}}\|_2 \le \sqrt{D}\|\boldsymbol{V}\bar{\boldsymbol{w}}\|_\infty \le \sqrt{D}(1 + 2\sqrt{D}) \le 3D. \tag{4.77}$$

By combining (4.15), (4.76) and (4.77), we finally obtain

$$\begin{aligned}
\int_{-\boldsymbol{V}^T\boldsymbol{p}+\bar{G}^*} g(\bar{\boldsymbol{w}})d\bar{\boldsymbol{w}} &= C(m,n)\int_{-\boldsymbol{V}^T\boldsymbol{p}+\bar{G}^*} \frac{1}{\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|^m}\left(1 + \frac{\|\bar{\boldsymbol{w}}\|^2}{\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|^2}\right)^{-(m+n)/2} d\bar{\boldsymbol{w}} \\
&> C(m,n)(2\sqrt{D})^{-m}(1 + 9D^2L^2/\epsilon^2)^{-(m+n)/2}\int_{-\boldsymbol{V}^T\boldsymbol{p}+\bar{G}^*} d\bar{\boldsymbol{w}} \\
&= C(m,n)(2\sqrt{D})^{-m}(1 + 9D^2L^2/\epsilon^2)^{-(m+n)/2}\,\mathrm{Vol}(-\boldsymbol{V}^T\boldsymbol{p} + \bar{G}^*) \\
&= C(m,n)(2\sqrt{D})^{-m}(1 + 9D^2L^2/\epsilon^2)^{-(m+n)/2}\,\mathrm{Vol}(\bar{G}^*),
\end{aligned}$$

where $C(m,n) = \Gamma((m+n)/2)/(\pi^{m/2}\Gamma(n/2))$ and where in the last equality we used the fact $\mathrm{Vol}(-\boldsymbol{V}^T\boldsymbol{p}+\bar{G}^*) = \mathrm{Vol}(\bar{G}^*)$ for any $\boldsymbol{p} \in \mathbb{R}^D$ (invariance of volumes under translations). The result follows from the assumption that G^* is non-degenerate, that is, $\mathrm{Vol}(\bar{G}^*) > 0$. $\square$

4.4 Summary

In this chapter, we considered the reduced problem (RP$\mathcal{X}$) associated to the constrained problem (P) for objective functions that have low effective dimensionality. The scope of this chapter was entirely theoretical in nature: our goal was to lower bound and quantify the probability of success of (RP$\mathcal{X}$). Similar to the analysis in Chapter 3, where we relied on the distribution of the least Euclidean norm minimizer $\boldsymbol{y}^*_2$ in $\mathcal{Y}$ space, for (RP$\mathcal{X}$), we based our analysis on the distribution of the minimizer $\boldsymbol{p} + \boldsymbol{A}\boldsymbol{y}^*_2$ in $\mathcal{X}$ space. We first established that $\boldsymbol{p}+\boldsymbol{A}\boldsymbol{y}_2^*$ can be expressed in terms of a multivariate t-distributed random variable, and using this reformulation, we obtained several lower bounds for the probability of success of (RP$\mathcal{X}$). In particular, we showed that $\mathbb{P}[$(RP$\mathcal{X}$) is successful$]$ is non-zero for any given $\boldsymbol{p} \in \mathcal{X}$. Furthermore, in the case of specific alignment of the effective subspace $\mathcal{T}$ with coordinate axes, it was possible to obtain lower bounds that are quantifiable, where one of the bounds (which is for $\boldsymbol{p} = \boldsymbol{0}$) exhibits an algebraic (and not exponential) dependence on D.

Moreover, we showed in Theorem 4.25 that, by enlarging the optimal set from exact to approximate minimizers and by additionally assuming Lipschitz continuity on f, we can lower bound the probability of ϵ-success of (RP$\mathcal{X}$) by a positive constant that does not depend on $\boldsymbol{p}$. This is an important result that will be used in Chapter 6 to prove global convergence of the algorithmic framework resulting from (RP$\mathcal{X}$). After the introduction and the analysis of this algorithmic framework in Chapter 6, we will conduct numerical experiments to test the effectiveness of the reduced formulation (RP$\mathcal{X}$).

We point out here that the above analysis assumes that $d \geq d_e$. In order to specify d in an(y) algorithm that uses (RP$\mathcal{X}$), one needs to know d_e or its upper bound, which are typically unknown in practice. See page 50 and the summary of Chapter 2, where we discuss this issue and propose possible solutions. Furthermore, recall that (RP) in addition needs to specify the parameter δ (the size of $\mathcal{Y}$ box) which requires the knowledge of not only d_e but also of μ, the distance from $\boldsymbol{p}$ to the closest minimizer. The advatange of (RP$\mathcal{X}$) in this regard is that we do not require the knowledge of μ or its upper bound to specify (RP$\mathcal{X}$) parameters.

The analysis in Section 4.3 and Theorem 4.25 leads us to the next chapter, where we further discuss ϵ-success of (RP$\mathcal{X}$), but this time, without assuming that f has low effective dimensionality. In the analysis, we only assume that f is Lipschitz continuous and based on properties of Gaussian matrices and conic integral geometry, we will attempt to bound the probability of ϵ-success of (RP$\mathcal{X}$).

Chapter 5

Global optimization of general functions using random embeddings

In this chapter, we investigate random embeddings technique applied to (P) for general objectives, which may not necessarily possess low effective structure. We tackle (P) for general objectives f with the same formulation (RP$\mathcal{X}$):

$$\begin{aligned} \min_{\boldsymbol{y}} \;& f(\boldsymbol{A}\boldsymbol{y}+\boldsymbol{p}) \\ \text{subject to } \;& \boldsymbol{A}\boldsymbol{y}+\boldsymbol{p} \in \mathcal{X}. \end{aligned} \tag{RP$\mathcal{X}$}$$

In the analysis of the present chapter, we only assume that f is Lipschitz continuous and for such f we are interested in estimating the probability that (RP$\mathcal{X}$) is ϵ-successful as defined in Definition 4.24.

For the objective function f that may not have a low effective structure, the set of global minimizers is no longer a collection of $(D - d_e)$-dimensional subspaces. Thus, techniques developed in the previous chapters cannot be straightforwardly applied to estimate the ϵ-success of (RP$\mathcal{X}$) for general functions. Therefore, we develop a new type of analysis based on connections between the reduced problem (RP$\mathcal{X}$) and the field of conic integral geometry, which we briefly discuss in Section 5.2. Relying on the tools from this field, in Section 5.3, we derive interpretable lower bounds (and their asymptotic behaviours) for the probability of ϵ-success of (RP$\mathcal{X}$). These bounds, unlike the bounds in the previous chapters, are valid for all embedding dimensions $d \geq 1$. At the end of the chapter, in order to understand the relative performance of (RP$\mathcal{X}$), we compare it with uniform sampling and determine asymptotic conditions under which (RP$\mathcal{X}$) is superior to uniform sampling in terms of lower bounds for the probability of ϵ-success.

We begin this chapter with a geometric description of (RP$\mathcal{X}$) applied to a general Lipschitz continuous objective function.

5.1 Geometric description

From the definitions of ϵ-success of (RP$\mathcal{X}$) and G_ϵ given in Definition 4.24 and (4.73), it follows that (RP$\mathcal{X}$) is ϵ-successful if and only if the intersection of the (affine) subspace $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ and G_ϵ is non-empty. This observation is equivalent to

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] = \mathbb{P}[\boldsymbol{p} + \text{range}(\boldsymbol{A}) \cap G_\epsilon \neq \varnothing]. \tag{5.1}$$

Now, suppose that f is Lipschitz continuous with Lipschitz constant L (i.e. f satisfies Assumption LipC) and let $\boldsymbol{x}^*$ be a global minimizer of f in $\mathcal{X}$. We consider a ball of radius ϵ/L around $\boldsymbol{x}^*$,

$$B_{\epsilon/L}(\boldsymbol{x}^*) = \{\boldsymbol{x} \in \mathcal{X} : \|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \leq \epsilon/L\}. \tag{5.2}$$

Note that if a point $\boldsymbol{x}$ lies in $B_{\epsilon/L}(\boldsymbol{x}^*)$ then it must also lie in G_ϵ due to the Lipschitz continuity property of f, namely

$$|f(\boldsymbol{x}) - f(\boldsymbol{x}^*)| \leq L\|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \leq L\frac{\epsilon}{L} = \epsilon. \tag{5.3}$$

Therefore, if the randomly drawn subspace $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ intersects $B_{\epsilon/L}(\boldsymbol{x}^*)$ then (RP$\mathcal{X}$) will be ϵ-successful. This observation together with (5.1) gives rise to the following relation:

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] \geq \mathbb{P}[\boldsymbol{p} + \text{range}(\boldsymbol{A}) \cap B_{\epsilon/L}(\boldsymbol{x}^*) \neq \varnothing]. \tag{5.4}$$

To estimate the latter probability, we first construct a set $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ containing the rays connecting $\boldsymbol{p}$ with points in $B_{\epsilon/L}(\boldsymbol{x}^*)$,

$$C_{\boldsymbol{p}}(\boldsymbol{x}^*) = \{\boldsymbol{p} + \theta(\boldsymbol{x} - \boldsymbol{p}) : \theta \geq 0, \boldsymbol{x} \in B_{\epsilon/L}(\boldsymbol{x}^*)\} \text{ for } \boldsymbol{p} \notin B_{\epsilon/L}(\boldsymbol{x}^*). \tag{5.5}$$

The illustration of $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ in Figure 5.1 and the nature of its formulation suggest that $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ is a convex cone that has been translated by $\boldsymbol{p}$. We can easily verify this fact by recalling the definition of a convex cone.

Definition 5.1. A convex set C is called a convex cone if for every $\boldsymbol{c} \in C$ and any non-negative scalar ρ, $\rho\boldsymbol{c} \in C$.

Remark 5.2. Note that, according to Definition 5.1, a d-dimensional linear subspace in $\mathbb{R}^D$ is a cone. Hence, range($\boldsymbol{A}$) is a cone.

We will later see that, based on (5.4) and the definition of $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$, we can lower bound the ϵ-success of (RP$\mathcal{X}$) by

$$\mathbb{P}[\boldsymbol{p} + \text{range}(\boldsymbol{A}) \cap C_{\boldsymbol{p}}(\boldsymbol{x}^*) \neq \{\boldsymbol{p}\}] \tag{5.6}$$

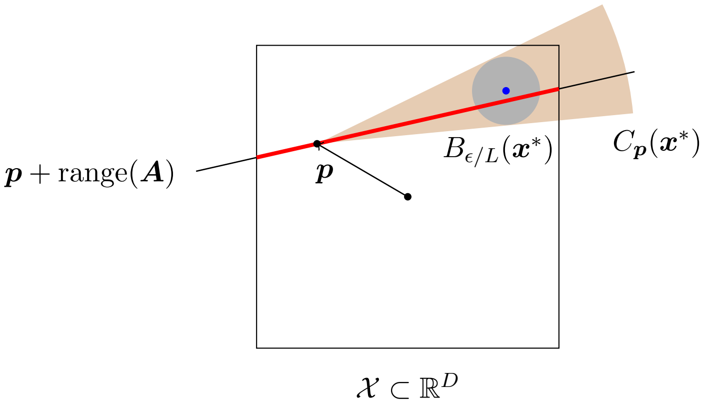

$\boldsymbol{p} + \text{range}(\boldsymbol{A})$
$\boldsymbol{p}$
$B_{\epsilon/L}(\boldsymbol{x}^*)$
$C_{\boldsymbol{p}}(\boldsymbol{x}^*)$
$\mathcal{X} \subset \mathbb{R}^D$

Figure 5.1: An abstract illustration of the embedding of an affine d-dimensional subspace $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ into $\mathbb{R}^D$. The red line represents the set of solutions along $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ that are contained in $\mathcal{X}$ and the blue dot represents a global minimizer $\boldsymbol{x}^*$ of (P). (RP$\mathcal{X}$) is ϵ-successful when the red line intersects $B_{\epsilon/L}(\boldsymbol{x}^*)$. We construct a cone $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ in such a way that the following condition holds: $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ intersects $B_{\epsilon/L}(\boldsymbol{x}^*)$ if and only if $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ and $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ share a ray.

— the probability of the event that translated cones $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ and $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ share a ray. It turns out that this probability has an exact quantifiable expression thanks to the field of conic integral geometry, where a main topic concerns the quantification/estimation of probabilities of a random cone (e.g., $\boldsymbol{p} + \text{range}(\boldsymbol{A})$) and a fixed cone (e.g., $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$) sharing a ray. In the next section, we recall key tools from conic integral geometry to help us estimate the ϵ-success of (RP$\mathcal{X}$).

5.2 A snapshot of conic integral geometry

A central question posed in conic integral geometry is the following:

> What is the probability that a randomly rotated convex cone shares a ray with a fixed convex cone?

The answer to this question is given by the conic kinematic formula [142].

Theorem 5.3 (Conic kinematic formula). *Let C and F be closed convex cones in $\mathbb{R}^D$ such that at most one of them is a linear subspace. Let $\boldsymbol{Q}$ be a $D \times D$ random orthogonal matrix drawn uniformly from the set of all $D \times D$ real orthogonal matrices. Then,*

$$\mathbb{P}[\boldsymbol{Q}F \cap C \neq \{\boldsymbol{0}\}] = \sum_{k=0}^{D} (1 + (-1)^{k+1}) \sum_{j=k}^{D} v_k(C) v_{D+k-j}(F), \tag{5.7}$$

where $v_k(C)$ denotes the kth intrinsic volume of cone C.

Proof. A proof can be found in [142, p. 261]. □

We would like to use the conic kinematic formula to estimate (5.6). The conic kinematic formula expresses the probability of the intersection of the two cones in terms of quantities known as conic intrinsic volumes. It is thus important to understand what these conic intrinsic volumes are and how to compute them.

5.2.1 Conic intrinsic volumes

Conic intrinsic volumes are commonly defined through the so called spherical Steiner formula (see [5]), which we will not mention here as it is beyond the scope of this work and it is not needed here. Instead, we will familiarize ourselves with the conic intrinsic volumes through their properties and specific examples. This is a short introductory review of conic intrinsic volumes; an interested reader is directed to [5, 6, 4, 104, 142] and the references therein.

For a closed convex cone C in $\mathbb{R}^D$, there are exactly $D+1$ conic intrinsic volumes: $v_0(C), v_1(C), \dots, v_D(C)$. Conic intrinsic volumes have many useful properties and below we summarize a few prominent ones. Given a closed convex cone $C \subseteq \mathbb{R}^D$, we have (see [6, Fact 5.5]):

(1) **Probability distribution**. The intrinsic volumes of the cone C are all nonnegative and sum up to 1, namely

$$\sum_{k=0}^{D} v_k(C) = 1 \text{ and } v_k(C) \geq 0 \text{ for } k = 0, 1, \dots, D. \tag{5.8}$$

In other words, they form a discrete probability distribution on $\{0, 1, \dots, D\}$.

(2) **Invariance under rotations**. Given any orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{D\times D}$, the intrinsic volumes of the rotated cone $\boldsymbol{Q}C$ and the original cone C are equal:

$$v_k(\boldsymbol{Q}C) = v_k(C) \text{ for } k = 0, 1, \dots, D. \tag{5.9}$$

(3) **Gauss-Bonnet formula**. If C is not a subspace, we have

$$\sum_{\substack{k=0 \\ k \text{ even}}}^{D} v_k(C) = \sum_{\substack{k=1 \\ k \text{ odd}}}^{D} v_k(C) = \frac{1}{2}. \tag{5.10}$$

The Gauss-Bonnet formula implies that $v_k(C) \leq 1/2$ for any k.

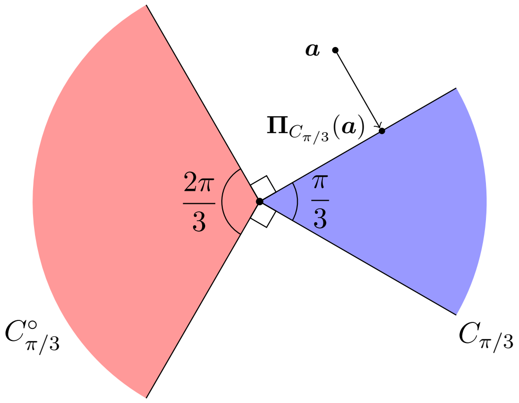


Figure 5.2: A depiction of the two-dimensional polyhedral cone $C_{\pi/3}$ in Example 5.6. The projection $\boldsymbol{\Pi}_{C_{\pi/3}}(\boldsymbol{a})$ of $\boldsymbol{a}$ onto $C_{\pi/3}$ falls onto the one-dimensional face of the cone.

Remark 5.4. One can think of conic intrinsic volumes as being cousins of the more familiar Euclidean intrinsic volumes. For a compact convex set K living in $\mathbb{R}^D$, Euclidean intrinsic volumes $v_0^E(K)$, $v_{D-1}^E(K)$ and $v_D^E(K)$ have familiar geometric interpretations: $v_0^E(K)$ — Euler characteristic, $2v_{D-1}^E(K)$ — surface area and $v_D^E(K)$ is the usual volume.

Remark 5.5. Another way of thinking about conic intrinsic volumes is through polyhedral cones — cones that can be generated by intersecting a finite number of halfspaces. If C is a polyhedral cone in $\mathbb{R}^D$, then the kth intrinsic volume of C is defined as follows (see [6, Definition 5.1])

$$v_k(C) := \mathbb{P}[\boldsymbol{\Pi}_C(\boldsymbol{a}) \text{ lies in the relative interior}^1 \text{of a } k\text{-dimensional face of } C]. \tag{5.11}$$

Here, $\boldsymbol{a}$ denotes the standard Gaussian vector[2] in $\mathbb{R}^D$ and $\boldsymbol{\Pi}_Y(\boldsymbol{x}) := \arg\min_{\boldsymbol{y}}\{\|\boldsymbol{x}-\boldsymbol{y}\| : \boldsymbol{y} \in Y\}$ denotes the Euclidean/orthogonal projection of $\boldsymbol{x}$ onto set Y, namely the vector in Y which is closest to $\boldsymbol{x}$.

Example 5.6. Let us consider a simple two-dimensional polyhedral cone $C_{\pi/3}$ illustrated in Figure 5.2 and let us calculate $v_0(C_{\pi/3})$, $v_1(C_{\pi/3})$ and $v_2(C_{\pi/3})$ using (5.11).

The cone $C_{\pi/3}$ has a single two-dimensional face (filled with blue), which is the interior of $C_{\pi/3}$. If a random vector $\boldsymbol{a}$ falls inside this face then $\boldsymbol{\Pi}_{C_{\pi/3}}(\boldsymbol{a}) = \boldsymbol{a}$ and, therefore,

$$v_2(C_{\pi/3}) = \mathbb{P}[\boldsymbol{a} \in C_{\pi/3}] = \frac{\pi/3}{2\pi} = \frac{1}{6}.$$

---

[1] The formal definition of the relative interior of set S is as follows: $\text{relint}(S) := \{\boldsymbol{x} \in S : \exists\delta > 0, B_\delta(\boldsymbol{x}) \cap \text{aff}(S) \subseteq S\}$, where the affine hull $\text{aff}(S)$ is the smallest affine set containing S. For example, the relative interior of a line segment $[A, B]$ living in $\mathbb{R}^2$ is (A, B); the relative interior of a two-dimensional square living in $\mathbb{R}^3$ is the square minus its boundary.

[2] A random vector for which each entry is an independent standard normal variable.

Let us now calculate $v_0(C_{\pi/3})$. Note that $C_{\pi/3}$ has only one zero-dimensional face, which is the origin. Note also that $\mathbf{\Pi}_{C_{\pi/3}}(\boldsymbol{a}) = \mathbf{0}$ if and only if $\boldsymbol{a} \in C^\circ_{\pi/3}$. Hence,

$$v_0(C_{\pi/3}) = \mathbb{P}[\boldsymbol{a} \in C^\circ_{\pi/3}] = \frac{2\pi/3}{2\pi} = \frac{1}{3}.$$

To calculate $v_1(C_{\pi/3})$, we simply use (5.8) to obtain

$$v_1(C_{\pi/3}) = 1 - v_0(C_{\pi/3}) - v_2(C_{\pi/3}) = \frac{1}{2}.$$

Example 5.7 (Linear subspace). The kth intrinsic volume of a d-dimensional linear subspace $\mathcal{L}_d$ in $\mathbb{R}^D$ is given by

$$v_k(\mathcal{L}_d) = \begin{cases} 1 & \text{if } k = d, \\ 0 & \text{otherwise.} \end{cases} \tag{5.12}$$

We already mentioned in Remark 5.2 that a d-dimensional linear subspace $\mathcal{L}_d$ is a cone. In fact, $\mathcal{L}_d$ is a polyhedral cone which has only one (d-dimensional) face. Therefore, the projection of any vector in $\mathbb{R}^D$ onto $\mathcal{L}_d$ will always lie on its (only) d-dimensional face. Hence, (5.12) follows from (5.11).

Example 5.8 (Circular cone). A circular cone is another important example; they have a number of applications in convex optimization (see, e.g., [16, Section 3] and [23, Section 4]). The circular cone of angle α in $\mathbb{R}^D$ is denoted by $\mathrm{Circ}_D(\alpha)$ and is defined as

$$\mathrm{Circ}_D(\alpha) := \{\boldsymbol{x} \in \mathbb{R}^D : x_1 \geq \|\boldsymbol{x}\| \cos(\alpha)\} \text{ for } 0 \leq \alpha \leq \pi/2.$$

The circular cone can be thought of as a collection of rays connecting the origin and some D-dimensional ball which does not contain the origin in its interior. The intrinsic volumes of $\mathrm{Circ}_D(\alpha)$ are given by the formulae (see [6, Appendix D.1]):

$$v_k(\mathrm{Circ}_D(\alpha)) = \frac{1}{2}\binom{(D-2)/2}{(k-1)/2} \sin^{k-1}(\alpha)\cos^{D-k-1}(\alpha) \tag{5.13}$$

for $k = 1, 2, \ldots, D-1$, where $\binom{i}{j}$ is the extension of the binomial coefficient to noninteger i and j through the gamma function,

$$\binom{i}{j} = \frac{\Gamma(i+1)}{\Gamma(j+1)\Gamma(i-j+1)}. \tag{5.14}$$

The 0th and Dth intrinsic volumes of the circular cone are given by (see [4, Ex. 4.4.8]):

$$v_0(\mathrm{Circ}_D(\alpha)) = \frac{D-1}{2}\binom{(D-2)/2}{-1/2} \int_0^{\pi/2-\alpha} \sin^{D-2}(x)dx, \tag{5.15}$$

$$v_D(\mathrm{Circ}_D(\alpha)) = \frac{D-1}{2}\binom{(D-2)/2}{(D-1)/2} \int_0^{\alpha} \sin^{D-2}(x)dx. \tag{5.16}$$

The following property of circular cones will be needed later.

Lemma 5.9. *Let* $\mathrm{Circ}_D(\alpha)$ *and* $\mathrm{Circ}_D(\beta)$ *be two circular cones with* $0 \le \alpha \le \beta \le \pi/2$. *Then,* $\mathrm{Circ}_D(\alpha) \subseteq \mathrm{Circ}_D(\beta)$.

Proof. Let $\boldsymbol{v}$ be any point in $\mathrm{Circ}_D(\alpha)$. By definition of $\mathrm{Circ}_D(\alpha)$, $v_1 \ge \|\boldsymbol{v}\|\cos(\alpha)$. Since $0 \le \alpha \le \beta \le \pi/2$, it follows that $v_1 \ge \|\boldsymbol{v}\|\cos(\beta)$, which by definition of $\mathrm{Circ}_D(\beta)$, implies that $\boldsymbol{v}$ must also lie in $\mathrm{Circ}_D(\beta)$. □

5.2.2 The Crofton formula

We now present a useful corollary of the conic kinematic formula. If one of the cones in Theorem 5.3 is given by a linear subspace then the conic kinematic formula reduces to the Crofton formula:

Corollary 5.10 (Crofton formula). *Let* C *be a closed convex cone in* $\mathbb{R}^D$ *and* $\mathcal{L}_d$ *be a* d-*dimensional linear subspace. Let* $\boldsymbol{Q}$ *be a* $D \times D$ *random orthogonal matrix drawn uniformly from the set of all* $D \times D$ *real orthogonal matrices. We have*

$$\mathbb{P}[\boldsymbol{Q}\mathcal{L}_d \cap C \neq \{\boldsymbol{0}\}] = \begin{cases} 2(v_{D-d+1}(C) + v_{D-d+3}(C) + \cdots + v_D(C)) & \textit{if } d \textit{ is odd,} \\ 2(v_{D-d+1}(C) + v_{D-d+3}(C) + \cdots + v_{D-1}(C)) & \textit{if } d \textit{ is even.} \end{cases} \tag{5.17}$$

The Crofton formula is easily derived from (5.7) using the fact that the kth intrinsic volume of linear subspace $\mathcal{L}_i$ is 1 if $i = k$ and 0 otherwise. The Crofton formula will be essential in estimating the probability of ϵ-success of (RP$\mathcal{X}$).

5.2.3 Connections to (RP$\mathcal{X}$)

Let us now mention the main implications of the above results for the analysis of ϵ-success of (RP$\mathcal{X}$). First, we note that if $\boldsymbol{p} \notin B_{\epsilon/L}(\boldsymbol{x}^*)$, then $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ defined in (5.5) is the circular cone $\mathrm{Circ}(\alpha^*_{\boldsymbol{p}})$ for $\alpha^*_{\boldsymbol{p}} = \arcsin(\epsilon/(L\|\boldsymbol{x}^* - \boldsymbol{p}\|))$ that has been rotated and then translated by $\boldsymbol{p}$. Therefore, in (5.6), we have an intersection of a random d-dimensional linear subspace and the rotated circular cone both translated by $\boldsymbol{p}$. We can translate these 'cones' back to the origin and then, using the Crofton formula, compute (5.6) exactly since the expressions for the conic intrinsic volumes of the circular cone $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ are known (see (5.13), (5.16) and (5.15)). The Crofton formula and (5.6) only differ in the formulation of a random linear subspace: in (5.17) a random linear subspace is given as $\boldsymbol{Q}\mathcal{L}_d$, whereas in (5.6) it is represented by range$(\boldsymbol{A})$. The following theorem states that these two representations are equivalent.

Theorem 5.11. *Let* $\boldsymbol{A} \in \mathbb{R}^{D\times d}$ *be a* $D \times d$ *Gaussian matrix. Let* $\boldsymbol{Q}$ *be a* $D \times D$ *random orthogonal matrix drawn uniformly from the set of all* $D \times D$ *real orthogonal matrices and let* $\mathcal{L}_d$ *be a d-dimensional linear subspace in* $\mathbb{R}^D$*. Then,*

$$\operatorname{range}(\boldsymbol{A}) \overset{law}{=} \boldsymbol{Q}\mathcal{L}_d. \tag{5.18}$$

Proof. See proof of [66, Theorem 1.2]. □

5.3 Bounding the probability of ϵ-success of the reduced problem (RP$\mathcal{X}$)

Relying on the tools of conic integral geometry and its connections to (RP$\mathcal{X}$), in this section, we derive two lower bounds for the ϵ-success of (RP$\mathcal{X}$): one that is dependent on $\boldsymbol{p}$ and one that holds uniformly over all $\boldsymbol{p} \in \mathcal{X}$.

Before we start, let us explicitly state the assumptions tacitly mentioned in the previous sections. Firstly, recall the definitions of $B_{\epsilon/L}(\boldsymbol{x}^*)$ and $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ in (5.2) and (5.5), respectively, and note that $B_{\epsilon/L}(\boldsymbol{x}^*)$ is a ball if and only if $\boldsymbol{x}^*$ is away from the boundary of $\mathcal{X}$ by distance at least ϵ/L.

Assumption 5.12. There exists a global minimizer $\boldsymbol{x^*}$ of (P) that satisfies $\operatorname{dist}(\boldsymbol{x}^*, \partial\mathcal{X}) \geq \epsilon/L$, where $\operatorname{dist}(\boldsymbol{x}^*, \partial\mathcal{X})$ denotes the minimal Euclidean distance from $\boldsymbol{x}^*$ to the boundary of $\mathcal{X}$.

Assumption 5.12 is reminiscent of the assumption in Theorem 4.25 that there exists an affine set of global minimizers G^* that is non-degenerate.

Next, we need to make sure that $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ is a well-defined circular cone when translated back to the origin. For that, in addition to Assumption 5.12 we require that $\boldsymbol{p}$ lies outside $B_{\epsilon/L}(\boldsymbol{x}^*)$. In the analysis below, we satisfy this condition[3] by assuming that $\boldsymbol{p} \notin G_\epsilon$. We exclude[4] cases $\boldsymbol{p} \notin G_\epsilon$ also since, for $\boldsymbol{p} \in G_\epsilon$, (RP$\mathcal{X}$) is ϵ-successful with probability 1.

We begin by formally proving the above mentioned fact that the probability of ϵ-success of (RP$\mathcal{X}$) is bounded below by (5.6).

Theorem 5.13. *Suppose that Assumptions LipC and 5.12 hold. Let* $\boldsymbol{A}$ *be a* $D\times d$ *Gaussian matrix and* ϵ*, a positive accuracy tolerance. Let* $\boldsymbol{p} \in \mathcal{X}\backslash G_\epsilon$ *be a given vector and let* $B_{\epsilon/L}(\boldsymbol{x}^*)$ *and* $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ *be defined in* (5.2) *and* (5.5)*, respectively. Then,*

$$\mathbb{P}[(\text{RP}\mathcal{X}) \textit{ is } \epsilon\textit{-successful}] \geq \mathbb{P}[\boldsymbol{p} + \operatorname{range}(\boldsymbol{A}) \cap C_{\boldsymbol{p}}(\boldsymbol{x}^*) \neq \{\boldsymbol{p}\}]. \tag{5.19}$$

---

[3] Note that, by (5.3), $\boldsymbol{p} \notin G_\epsilon \Rightarrow \boldsymbol{p} \notin B_{\epsilon/L}(\boldsymbol{x}^*)$.

[4] It is worth mentioning that the assumption $\boldsymbol{p} \notin G_\epsilon$ is equivalent to the assumption $\boldsymbol{p}_\top \neq \boldsymbol{x}^*_\top$ in the analysis of Chapters 3 and 4.

Proof. From (5.1) and (5.4), we have

$$\begin{aligned}\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] &= \mathbb{P}[\boldsymbol{p} + \mathrm{range}(\boldsymbol{A}) \cap G_\epsilon \neq \varnothing] \\ &\geq \mathbb{P}[\boldsymbol{p} + \mathrm{range}(\boldsymbol{A}) \cap B_{\epsilon/L}(\boldsymbol{x}^*) \neq \varnothing].\end{aligned}$$

The result follows from the fact that the event $\{\boldsymbol{p} + \mathrm{range}(\boldsymbol{A}) \cap C_{\boldsymbol{p}}(\boldsymbol{x}^*) \neq \{\boldsymbol{p}\}\}$ is a subset of the event $\{\boldsymbol{p} + \mathrm{range}(\boldsymbol{A}) \cap B_{\epsilon/L}(\boldsymbol{x}^*) \neq \varnothing\}$. We prove this fact below.

Suppose that the event $\{\boldsymbol{p} + \mathrm{range}(\boldsymbol{A}) \cap C_{\boldsymbol{p}}(\boldsymbol{x}^*) \neq \{\boldsymbol{p}\}\}$ occurs. Then, there exists a point $\boldsymbol{x}' \neq \boldsymbol{p}$ in $\boldsymbol{p} + \mathrm{range}(\boldsymbol{A}) \cap C_{\boldsymbol{p}}(\boldsymbol{x}^*)$. Define $R = \{\boldsymbol{p} + \theta(\boldsymbol{x}' - \boldsymbol{p}) : \theta \geq 0\}$ and note that $R \subset \boldsymbol{p} + \mathrm{range}(\boldsymbol{A})$. Now, since $\boldsymbol{x}' \in C_{\boldsymbol{p}}(\boldsymbol{x}^*)$, by definition of $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ there exists $\tilde{\boldsymbol{x}} \in B_{\epsilon/L}(\boldsymbol{x}^*)$ and $\tilde{\theta} > 0$ such that $\boldsymbol{x}' = \boldsymbol{p} + \tilde{\theta}(\tilde{\boldsymbol{x}} - \boldsymbol{p})$. We express $\tilde{\boldsymbol{x}}$ in terms of $\boldsymbol{x}'$: $\tilde{\boldsymbol{x}} = \boldsymbol{p} + \theta'(\boldsymbol{x}' - \boldsymbol{p})$, where $\theta' = 1/\tilde{\theta} > 0$. By definition of R, $\tilde{\boldsymbol{x}} \in R$ and, thus, $\tilde{\boldsymbol{x}}$ also lies in $\boldsymbol{p} + \mathrm{range}(\boldsymbol{A})$. This proves that the set $\{\boldsymbol{p} + \mathrm{range}(\boldsymbol{A}) \cap B_{\epsilon/L}(\boldsymbol{x}^*)\}$ is non-empty. □

The sets $\boldsymbol{p} + \mathrm{range}(\boldsymbol{A})$ and $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ are two translated cones. To be able to apply the Crofton formula to (5.19), we need to translate these sets back to the origin and rotate them appropriately so that $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ becomes a circular cone. We transform (5.19) into a 'Crofton formula'-suitable form in the following corollary.

Corollary 5.14. *Suppose that the conditions in Theorem 5.13 hold. Let* $\boldsymbol{Q}$ *be a* $D \times D$ *random orthogonal matrix drawn uniformly from the set of all* $D\times D$ *real orthogonal matrices and* $\mathcal{L}_d$ *be a d-dimensional linear subspace in* $\mathbb{R}^D$. *Let* $\mathrm{Circ}(\alpha^*_{\boldsymbol{p}})$ *be the circular cone with* $\alpha^*_{\boldsymbol{p}} = \arcsin(\epsilon/(L\|\boldsymbol{x}^* - \boldsymbol{p}\|))$. *Then,*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \textit{ is } \epsilon\textit{-successful}] \geq \mathbb{P}[\boldsymbol{Q}\mathcal{L}_d \cap \mathrm{Circ}(\alpha^*_{\boldsymbol{p}}) \neq \{\boldsymbol{0}\}]. \tag{5.20}$$

Proof. As mentioned earlier, by definition, $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ is a rotated and translated (by $\boldsymbol{p}$) circular cone $\mathrm{Circ}(\alpha^*_{\boldsymbol{p}})$. That is, there exists a $D \times D$ orthogonal matrix $\boldsymbol{S}$ such that $C_{\boldsymbol{p}}(\boldsymbol{x}^*) = \boldsymbol{p} + \boldsymbol{S}\,\mathrm{Circ}(\alpha^*_{\boldsymbol{p}})$. Then, Theorem 5.13 implies

$$\begin{aligned}\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] &\geq \mathbb{P}[\boldsymbol{p} + \mathrm{range}(\boldsymbol{A}) \cap \boldsymbol{p} + \boldsymbol{S}\,\mathrm{Circ}(\alpha^*_{\boldsymbol{p}}) \neq \{\boldsymbol{p}\}] \\ &= \mathbb{P}[\mathrm{range}(\boldsymbol{A}) \cap \boldsymbol{S}\,\mathrm{Circ}(\alpha^*_{\boldsymbol{p}}) \neq \{\boldsymbol{0}\}] \\ &= \mathbb{P}[\boldsymbol{S}^T\,\mathrm{range}(\boldsymbol{A}) \cap \mathrm{Circ}(\alpha^*_{\boldsymbol{p}}) \neq \{\boldsymbol{0}\}] \qquad (5.21) \\ &= \mathbb{P}[\mathrm{range}(\boldsymbol{A}) \cap \mathrm{Circ}(\alpha^*_{\boldsymbol{p}}) \neq \{\boldsymbol{0}\}] \\ &= \mathbb{P}[\boldsymbol{Q}\mathcal{L}_d \cap \mathrm{Circ}(\alpha^*_{\boldsymbol{p}}) \neq \{\boldsymbol{0}\}],\end{aligned}$$

where the penultimate equality follows the orthogonal invariance of Gaussian matrices and where the last equality follows from Theorem 5.11. □

Corollary 5.14 now allows us to use the Crofton formula to quantify the lower bound in (5.20). In the next theorem, we derive our first lower bound, that is dependent on the location of $\boldsymbol{p}$ in $\mathcal{X}$.

Theorem 5.15 (A lower bound). *Suppose that Assumptions LipC and 5.12 hold. Let* $\boldsymbol{A}$ *be a* $D \times d$ *Gaussian matrix and* $\epsilon > 0$*, an accuracy tolerance. Let* $\boldsymbol{p} \in \mathcal{X} \setminus G_\epsilon$ *be a given vector and let* $r_{\boldsymbol{p}} := \epsilon/(L\|\boldsymbol{x}^* - \boldsymbol{p}\|)$*. Then,*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \textit{ is } \epsilon\textit{-successful}] \geq \tau(r_{\boldsymbol{p}}), \tag{5.22}$$

where the function $\tau(r)$ *for* $0 < r < 1$ *is defined as*

$$\tau(r) := \begin{cases} (D-1) \cdot \binom{\frac{D-2}{2}}{\frac{D-1}{2}} \int_0^{\arcsin(r)} \sin^{D-2}(x)dx & \textit{if } d = 1, \\ \binom{\frac{D-2}{2}}{\frac{D-d}{2}} r^{D-d}(1-r^2)^{\frac{d-2}{2}} & \textit{if } 1 < d < D. \end{cases} \tag{5.23}$$

Here, $\binom{i}{j}$ *denotes the general binomial coefficient defined in* (5.14).

Proof. Let $\alpha^*_{\boldsymbol{p}} = \arcsin(r_{\boldsymbol{p}})$ and let C denote $\mathrm{Circ}(\alpha^*_{\boldsymbol{p}})$ for notational convenience. First, note that by (5.13) and (5.16), $\tau(r) = 2v_{D-d+1}(\mathrm{Circ}(\arcsin(r)))$. Thus, all we need to show is that $\mathbb{P}[(\mathrm{RP}\mathcal{X})$ is ϵ-successful$]$ is lower bounded by $2v_{D-d+1}(C)$.

By (5.20) and the Crofton formula (5.17), we have

$$\begin{aligned} \mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] &\geq \begin{cases} 2(v_{D-d+1}(C) + v_{D-d+3}(C) + \cdots + v_D(C)) & \text{if } d \text{ is odd,} \\ 2(v_{D-d+1}(C) + v_{D-d+3}(C) + \cdots + v_{D-1}(C)) & \text{if } d \text{ is even.} \end{cases} \\ &\geq 2v_{D-d+1}(C), \end{aligned} \tag{5.24}$$

where the inequality follows from the fact that $v_k(C)$'s are all nonnegative (see (5.8)). ☐

One may be wondering why we choose to bound the ϵ-success of $(\mathrm{RP}\mathcal{X})$ in (5.24) by the multiple of $v_{D-d+1}(C)$ in particular, whereas we could have chosen any other intrinsic volume or kept the sum altogether. Our rationale for such a choice of the lower bound is underpinned by the following observation: using the formulae (5.13) and (5.16) for the intrinsic volumes, one can verify that $v_{D-d+i}(C)/v_{D-d+1}(C) = O(D^{(1-i)/2})$ for $i = 1, 2, \ldots, d$ as $D \to \infty$ with other parameters kept fixed[5]. Hence,

$$v_{D-d+1}(C) + v_{D-d+3}(C) + \cdots = v_{D-d+1}(C) \cdot (1 + O(1/D)).$$

Therefore, approximating the sum by its leading term $v_{D-d+1}(C)$ should produce a reasonably tight bound for large values of D.

Next, we derive a lower bound that holds uniformly for all $\boldsymbol{p} \in \mathcal{X}$.

---

[5] The term $v_{D-d+1}(C)$ is dominant also in the scenario when $\|\boldsymbol{x}^* - \boldsymbol{p}\| \to \infty$ as $D \to \infty$ with other parameters fixed. In this case, $v_{D-d+i}(C)/v_{D-d+1}(C) = O((r_{\boldsymbol{p}}/\sqrt{D})^{i-1})$ for $i = 1, 2, \ldots, d$ as $D \to \infty$.

Theorem 5.16 (A uniform lower bound). *Suppose that Assumptions LipC and 5.12 hold. Let* $\boldsymbol{A}$ *be a* $D \times d$ *Gaussian matrix and* ϵ*, a positive accuracy tolerance. Let* $r_{min} := \epsilon/(2L\sqrt{D})$*. For all* $\boldsymbol{p} \in \mathcal{X}$*, we have*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \textit{ is } \epsilon\textit{-successful}] \geq \tau(r_{min}), \tag{5.25}$$

where $\tau(\cdot)$ *is defined in* (5.23)*.*

Proof. Let $\boldsymbol{x}^*$ be a global minimizer that satisfies Assumption 5.12, let $r_{\boldsymbol{p}}$ be defined in Theorem 5.15 and let $\alpha^*_{\boldsymbol{p}} = \arcsin(r_{\boldsymbol{p}})$. We consider cases $\boldsymbol{p} \in \mathcal{X} \setminus G_\epsilon$ and $\boldsymbol{p} \in G_\epsilon$ separately.

First, let $\boldsymbol{p}$ be any point in $\mathcal{X} \setminus G_\epsilon$. Note that since $\boldsymbol{x}^*, \boldsymbol{p} \in [-1,1]^D$, $\|\boldsymbol{x}^* - \boldsymbol{p}\| \leq 2\sqrt{D}$ for all $\boldsymbol{p} \in \mathcal{X}$. Therefore, for any $\boldsymbol{p} \in \mathcal{X} \setminus G_\epsilon$, the following holds

$$\begin{aligned} r_{\boldsymbol{p}} &\geq \frac{\epsilon}{2L\sqrt{D}} = r_{min}, \\ \alpha^*_{\boldsymbol{p}} &\geq \arcsin(r_{min}) := \alpha^*_{min}. \end{aligned} \tag{5.26}$$

Now, define $C_{min} := \mathrm{Circ}(\alpha^*_{min})$. By (5.26) and Lemma 5.9, it follows that $C_{min} \subseteq \mathrm{Circ}(\alpha^*_{\boldsymbol{p}})$. Using Corollary 5.14, we then obtain

$$\begin{aligned} \mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] &\geq \mathbb{P}[\boldsymbol{Q}\mathcal{L}_d \cap \mathrm{Circ}(\alpha^*_{\boldsymbol{p}}) \neq \{\boldsymbol{0}\}] \\ &\geq \mathbb{P}[\boldsymbol{Q}\mathcal{L}_d \cap C_{min} \neq \{\boldsymbol{0}\}] \\ &\geq 2v_{D-d+1}(C_{min}), \end{aligned} \tag{5.27}$$

where the last inequality follows from the same line of argument as in (5.24). Using (5.13) and (5.16), it is easy to verify that $2v_{D-d+1}(C_{min}) = \tau(r_{min})$. We have shown (5.25) for $\boldsymbol{p} \in \mathcal{X} \setminus G_\epsilon$.

For $\boldsymbol{p} \in G_\epsilon$, (5.25) holds trivially, since if $\boldsymbol{p} \in G_\epsilon$, $(\mathrm{RP}\mathcal{X})$ is ϵ-successful with probability 1. As a sanity check, $1 \geq 2v_{D-d+1}(C_{min}) = \tau(r_{min})$ where the inequality is implied by the Gauss-Bonnet formula (5.10). ☐

Theorem 5.16 provides a lower bound valid for even the worst possible scenario: when $\boldsymbol{x}^*$ and $\boldsymbol{p}$ are near the opposite corners of $\mathcal{X}$.

Both $\tau(r_{\boldsymbol{p}})$ and $\tau(r_{min})$ are defined through $\tau(r)$, whose formula is not easily interpretable. To better understand the dependence of both lower bounds on the parameters of the problem, in the next section, we analyse the behaviour of $\tau(r)$, and consequently of $\tau(r_{\boldsymbol{p}})$ and $\tau(r_{min})$, in the asymptotic regime.

5.3.1 Asymptotic expansions

We establish asymptotic behaviours of $\tau(r_{\boldsymbol{p}})$ and $\tau(r_{min})$ for large D. The other parameters are kept fixed except for $\|\boldsymbol{x}^* - \boldsymbol{p}\|$ in $\tau(r_{\boldsymbol{p}})$, which we also allow to increase with D. Before we begin, we first need to establish the following lemma.

Lemma 5.17. *Let $0 < \alpha < \pi/2$ be either a fixed angle or a function of D that tends to 0 as $D \to \infty$. Then, as $D \to \infty$,*

$$\int_0^\alpha \sin^D(x)dx = \frac{1}{D}\frac{\sin^{D+1}(\alpha)}{\cos(\alpha)} + O\left(\frac{\sin^{D+1}(\alpha)}{D^2}\right). \tag{5.28}$$

Proof. We write

$$\int_0^\alpha \sin^D(x)dx = \int_0^\alpha \frac{\sin(x)}{D\cos(x)} \cdot (D\cos(x)\sin^{D-1}(x))dx. \tag{5.29}$$

Integration by parts with $u = \sin(x)/(D\cos(x))$ and $dv = D\cos(x)\sin^{D-1}(x)dx$ yields

$$\int_0^\alpha \sin^D(x)dx = \frac{\sin^{D+1}(\alpha)}{D\cos(\alpha)} - \frac{1}{D}\int_0^\alpha \frac{\sin^D(x)}{\cos^2(x)}dx. \tag{5.30}$$

Let I denote $\int_0^\alpha \frac{\sin^D(x)}{\cos^2(x)}dx$. It remains to show that $I = O(\sin^{D+1}(\alpha)/D)$. We express I as

$$\int_0^\alpha \frac{\sin(x)}{D\cos^3(x)} \cdot (D\cos(x)\sin^{D-1}(x))dx. \tag{5.31}$$

We integrate I by parts with $u = \sin(x)/(D\cos^3(x))$ and $dv = D\cos(x)\sin^{D-1}(x)dx$ to obtain

$$I = \frac{1}{D}\frac{\sin^{D+1}(\alpha)}{\cos^3(\alpha)} - \frac{1}{D}\int_0^\alpha \frac{1+2\sin^2(x)}{\cos^4(x)}\sin^D(x)dx \tag{5.32}$$

Since the latter integral is positive, we have

$$I \leq \frac{1}{\cos^3(\alpha)} \cdot \frac{\sin^{D+1}(\alpha)}{D}. \tag{5.33}$$

Since I is positive for any $0 < \alpha < \pi/2$, (5.33) implies that $I = O(\sin^{D+1}(\alpha)/D)$. ☐

We establish the asymptotic behaviours of $\tau(r_{\boldsymbol{p}})$ and $\tau(r_{min})$ by analysing the asymptotics of $\tau(r)$ defined in (5.23) and later substituting $r_{\boldsymbol{p}}$ and r_{min} for r in $\tau(r)$.

Theorem 5.18. *Let $\tau(r)$ be defined in* (5.23)*. Let d be fixed and let r be either fixed or tend to zero as $D \to \infty$. Then,*

$$\tau(r) = \Theta\left(D^{\frac{d-2}{2}} r^{D-d}\right) \text{ as } D \to \infty, \tag{5.34}$$

and the constants in $\Theta(\cdot)$ are independent of D.

Proof. We prove (5.34) for $d = 1$ and $1 < d < D$ separately.

First, assume that $d > 1$. By definition of $\tau(r)$, we have

$$\tau(r) = \binom{\frac{D-2}{2}}{\frac{D-d}{2}} r^{D-d}(1-r^2)^{\frac{d-2}{2}}. \tag{5.35}$$

Let us first determine the asymptotic behaviour of the binomial coefficient. Using the fact that $\Gamma(z+a)/\Gamma(z+b) = \Theta(z^{a-b})$ for large z (see, e.g., [155]), we obtain

$$\binom{\frac{D-2}{2}}{\frac{D-d}{2}} = \frac{\Gamma(\frac{D}{2})}{\Gamma(\frac{D-d+2}{2})\Gamma(\frac{d}{2})} = \frac{\Gamma(\frac{D-d+2}{2} + \frac{d-2}{2})}{\Gamma(\frac{D-d+2}{2})\Gamma(\frac{d}{2})} = \Theta\left(\left(\frac{D-d+2}{2}\right)^{\frac{d-2}{2}}\right) = \Theta\left(D^{\frac{d-2}{2}}\right). \tag{5.36}$$

To obtain[6] (5.34), we substitute (5.36) into (5.35). Note that $(1-r^2)^{\frac{d-2}{2}}$ is bounded above and bounded away from zero by constants independent of D; thus, it can be absorbed into the constants of Θ.

Let us now prove (5.34) for $d = 1$. We have

$$\tau(r) = (D-1) \cdot \binom{\frac{D-2}{2}}{\frac{D-1}{2}} \int_0^{\arcsin(r)} \sin^{D-2}(x)dx, \tag{5.37}$$

where, by (5.36),

$$\binom{\frac{D-2}{2}}{\frac{D-1}{2}} = \Theta\left(D^{-\frac{1}{2}}\right) \tag{5.38}$$

and, by Lemma 5.17,

$$\int_0^{\arcsin(r)} \sin^{D-2}(x)dx = \Theta\left(\frac{1}{D-1}\frac{r^{D-1}}{\sqrt{1-r^2}}\right). \tag{5.39}$$

By substituting (5.38) and (5.39) into (5.37), we obtain (5.34) for $d = 1$. For similar reasons as stated above, we can relegate the term $1/\sqrt{1-r^2}$ in (5.39) into the constants of Θ. □

Now, to obtain the asymptotics for $\tau(r_{\boldsymbol{p}})$, we simply apply Theorem 5.18 for $r = r_{\boldsymbol{p}} = \epsilon/(L\|\boldsymbol{x}^* - \boldsymbol{p}\|)$.

Corollary 5.19. *Let the conditions in Theorem 5.15 hold. Let parameters d, ϵ, L be fixed and let $\|\boldsymbol{x}^* - \boldsymbol{p}\|$ be either fixed or tend to infinity as $D \to \infty$. Then,*

$$\tau(r_{\boldsymbol{p}}) = \Theta\left(D^{\frac{d-2}{2}}\left(\frac{\epsilon}{L\|\boldsymbol{x}^* - \boldsymbol{p}\|}\right)^{D-d}\right) \quad \text{as } D \to \infty, \tag{5.40}$$

and the constants in $\Theta(\cdot)$ are independent of D.

Proof. Note that $r_{\boldsymbol{p}} = \epsilon/(L\|\boldsymbol{x}^* - \boldsymbol{p}\|)$ is either fixed or tends to zero as $D \to \infty$. Then, the result follows from Theorem 5.18. □

Corollary 5.19 shows that for any $\boldsymbol{p}$ outside G_ϵ, the lower bound $\tau(r_{\boldsymbol{p}})$ decreases exponentially with D. Note that this decrease is slower for larger values of d or $\boldsymbol{p}$ closer to $\boldsymbol{x}^*$. To obtain the asymptotics for the uniform lower bound $\tau(r_{min})$, we apply Theorem 5.18 for $r = r_{min} = \epsilon/(2L\sqrt{D})$.

---

[6] Here, we have also used the fact that if functions $f(x)$, $\bar{f}(x)$, $g(x)$ and $\bar{g}(x)$ satisfy $f(x) = \Theta(g(x))$ and $\bar{f}(x) = \Theta(\bar{g}(x))$ (as $x \to \infty$), then $f(x)\bar{f}(x) = \Theta(g(x)\bar{g}(x))$.

Corollary 5.20. *Let conditions in Theorem 5.16 hold. Let parameters d, ϵ, L be fixed. Then,*

$$\tau(r_{min}) = \Theta\left(D^{-\frac{D}{2}+d-1}\left(\frac{\epsilon}{2L}\right)^{D-d}\right) \text{ as } D \to \infty, \tag{5.41}$$

and the constants in $\Theta(\cdot)$ are independent of D.

Proof. Note that $r_{min} = \epsilon/(2L\sqrt{D})$ tends to zero as $D \to \infty$. Then, the result follows from Theorem 5.18. $\square$

5.3.2 Comparing (RP$\mathcal{X}$) to a simple random search

Using the above estimates for the probability of ϵ-success of (RP$\mathcal{X}$), we now would like to compare (RP$\mathcal{X}$) to a simple random search method to understand the relative performance of (RP$\mathcal{X}$) for general functions. As a baseline for comparison, we use Uniform Sampling (US). We start off with the derivation of a lower bound for the probability of ϵ-success of US and the computation of its asymptotics.

Note that if a uniformly sampled point falls inside $B_{\epsilon/L}(\boldsymbol{x}^*)$ then US is ϵ-successful. This implies that

$$\mathbb{P}[\text{US is } \epsilon\text{-successful}] \geq \frac{\text{Vol}(B_{\epsilon/L}(\boldsymbol{x}^*))}{\text{Vol}(\mathcal{X})} = \frac{\pi^{D/2}}{2^D\Gamma(\frac{D}{2}+1)}\left(\frac{\epsilon}{L}\right)^D := \tau_{us}, \tag{5.42}$$

where we have used the fact that $\text{Vol}(B_{\epsilon/L}(\boldsymbol{x}^*)) = \frac{\pi^{D/2}}{\Gamma(\frac{D}{2}+1)}\left(\frac{\epsilon}{L}\right)^D$ (see [121, Equation 5.19.4]) and that $\text{Vol}(\mathcal{X}) = 2^D$.

Using Stirling's approximation, it is straightforward to establish the asymptotic behaviour of the lower bound τ_{us}.

Lemma 5.21. *Let τ_{us} be defined in* (5.42) *and let ϵ and L be fixed. Then,*

$$\tau_{us} = \Theta\left(D^{-\frac{D}{2}-\frac{1}{2}}\left(\frac{\pi e}{2}\right)^{\frac{D}{2}}\left(\frac{\epsilon}{L}\right)^D\right) \text{ as } D \to \infty. \tag{5.43}$$

Proof. By Stirling's approximation (see [121, Equation 5.11.7]),

$$\Gamma\left(\frac{D}{2}+1\right) = \Theta\left(e^{-\frac{D}{2}}\left(\frac{D}{2}\right)^{\frac{D+1}{2}}\right) \text{ as } D \to \infty. \tag{5.44}$$

By substituting (5.44) into (5.42), we obtain the desired result. $\square$

Let us now compare the lower bound τ_{us} of US to the lower bound $\tau(r_{\boldsymbol{p}})$ for (RP$\mathcal{X}$). It is clear from the analysis of $\tau(r_{\boldsymbol{p}})$ in Section 5.3.1 that the probability of ϵ-success of (RP$\mathcal{X}$) is higher if $\boldsymbol{p}$ is closer to the set of global minimizers. In the next theorem, we determine a threshold distance Δ_0 between $\boldsymbol{p}$ and a global minimizer $\boldsymbol{x}^*$ such that $\tau(r_{\boldsymbol{p}})$

and τ_{us} are approximately equal to each other. This would tell us how close $\boldsymbol{p}$ should be to $\boldsymbol{x}^*$ for (RP$\mathcal{X}$) to have a larger lower bound for the probability of success than that of US. The analysis is done in the asymptotic regime.

Theorem 5.22. *Suppose that Assumptions LipC and 5.12 hold. Let $\tau(r_{\boldsymbol{p}})$ and τ_{us} be defined in Theorem 5.15 and (5.42), respectively. Let ϵ, L, d be fixed and let $\Delta_0 = \sqrt{\frac{2D}{\pi e}}$. Then,*

a) If $\lim_{D\to\infty} \dfrac{\Delta_0}{\|\boldsymbol{x}^* - \boldsymbol{p}\|} = \psi > 1$, then $\tau(r_{\boldsymbol{p}})/\tau_{us} \to \infty$ as $D \to \infty$.

b) If $\lim_{D\to\infty} \dfrac{\Delta_0}{\|\boldsymbol{x}^* - \boldsymbol{p}\|} = \psi < 1$, then $\tau(r_{\boldsymbol{p}})/\tau_{us} \to 0$ as $D \to \infty$.

Proof. From (5.40) and (5.43), we have

$$\frac{\tau(r_{\boldsymbol{p}})}{\tau_{us}} = \frac{\Theta\left(D^{\frac{d-2}{2}}\left(\frac{\epsilon}{L\|\boldsymbol{x}^*-\boldsymbol{p}\|}\right)^{D-d}\right)}{\Theta\left(\frac{1}{\sqrt{D}}\left(\frac{\pi e}{2D}\right)^{\frac{D}{2}}\left(\frac{\epsilon}{L}\right)^{D}\right)} \overset{7}{=} \Theta\left(\left(\frac{\epsilon}{L}\right)^{-d}\left(\frac{2}{\pi e}\right)^{D/2} D^{\frac{D+d-1}{2}}\|\boldsymbol{x}^*-\boldsymbol{p}\|^{d-D}\right)$$
$$= \Theta\left(\left(\underbrace{\left[\frac{\sqrt{2D/\pi e}}{\|\boldsymbol{x}^*-\boldsymbol{p}\|}\right]}_{=\Delta_0/\|\boldsymbol{x}^*-\boldsymbol{p}\|}\cdot D^{\frac{2d-1}{2(D-d)}}\right)^{D-d}\right), \tag{5.45}$$

Note that in the second line there is a term $\left(\frac{\epsilon}{L}\right)^{-d}\left(\frac{2}{\pi e}\right)^{d/2}$ missing inside Θ, which we removed as it is independent of D. Now, by definition of Θ, (5.45) implies that there exist positive constants M_1 and M_2 and a positive integer D_0 such that

$$M_1\left(\frac{\Delta_0}{\|\boldsymbol{x}^*-\boldsymbol{p}\|}D^{\frac{2d-1}{2(D-d)}}\right)^{D-d} \le \frac{\tau(r_{\boldsymbol{p}})}{\tau_{us}} \le M_2\left(\frac{\Delta_0}{\|\boldsymbol{x}^*-\boldsymbol{p}\|}D^{\frac{2d-1}{2(D-d)}}\right)^{D-d} \quad \text{for all } D \ge D_0. \tag{5.46}$$

Note that $D^{\frac{2d-1}{2(D-d)}} \to 1$ as $D \to \infty$. Hence, if $\Delta_0/\|\boldsymbol{x}^*-\boldsymbol{p}\| \to \psi > 1$ then both lower and upper bounds in (5.46) tend to infinity implying that $\tau(r_{\boldsymbol{p}})/\tau_{us} \to \infty$. On the other hand, if $\Delta_0/\|\boldsymbol{x}^*-\boldsymbol{p}\| \to \psi < 1$ then both lower and upper bounds in (5.46) tend to zero implying that $\tau(r_{\boldsymbol{p}})/\tau_{us} \to 0$. □

Theorem 5.22 tells us that the distance between $\boldsymbol{p}$ and $\boldsymbol{x}^*$ (in the asymptotic setting) must be no greater than $\Delta_0 \approx 0.48\sqrt{D}$ for $\tau(r_{\boldsymbol{p}})$ to be larger than τ_{us}. Note that, since the distance between the origin and a corner of $\mathcal{X}$ is equal to $\sqrt{D}$ ($> 0.48\sqrt{D}$), there is no point $\boldsymbol{p}$ such that the ball of radius Δ_0 centred at $\boldsymbol{p}$ covers all points in $\mathcal{X}$. In other words, for any $\boldsymbol{p}$ in $\mathcal{X}$, there always exists $\boldsymbol{x}^*$ for which $\tau(r_{\boldsymbol{p}})$ is smaller than τ_{us}. On the other hand, if $\boldsymbol{p} = \boldsymbol{0}$ and $\boldsymbol{x}^*$ is close to the origin then $\tau(r_{\boldsymbol{p}}) > \tau_{us}$. Note also that Δ_0 has no

---

[7] Here, we use the fact that if functions $f(x)$, $\bar{f}(x)$, $g(x)$ and $\bar{g}(x)$ satisfy $f(x) = \Theta(g(x))$ and $\bar{f}(x) = \Theta(\bar{g}(x))$ (as $x \to \infty$), then $f(x)/\bar{f}(x) = \Theta(g(x)/\bar{g}(x))$.

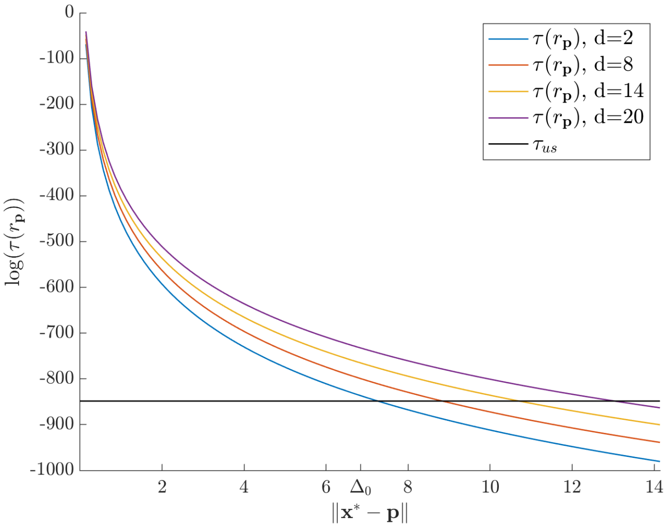


Figure 5.3: A plot of $\tau(r_{\boldsymbol{p}})$ versus $\|\boldsymbol{x}^* - \boldsymbol{p}\|$ for different values of the subspace embedding dimension d. The lower bound τ_{us} of US does not depend on $\|\boldsymbol{x}^* - \boldsymbol{p}\|$ and, thus, it is displayed as a straight horizontal line.

dependence on the embedding subspace dimension d. This is due to the asymptotic nature of the analysis: in (5.46), we see that both inequalities depend on d, but the dependence diminishes as $D \to \infty$ since d is kept fixed. Although the asymptotic analysis shows no significant dependence on the subspace dimension, numerical experiments show that the value of d has a notable effect on success of (RP$\mathcal{X}$). In Figure 5.3, we plot $\tau(r_{\boldsymbol{p}})$ as a function of $\|\boldsymbol{x}^* - \boldsymbol{p}\|$ for different values of d with D fixed at 200. The lower bound τ_{us} of US is represented by a black horizontal line. We see that, for larger d, $\tau(r_{\boldsymbol{p}})$ decreases at a slower rate and has greater threshold distance before becoming smaller than τ_{us}.

Remark 5.23. An important distinction must be made between the implications of the ϵ-success of (RP$\mathcal{X}$) and the ϵ-success of US in solving the original problem (P). Note that the ϵ-success of US means that US has sampled a point that lies in G_ϵ, which in turn implies that US has successfully (approximately) solved (P). This is not the case for (RP$\mathcal{X}$). Recall that ϵ-success of (RP$\mathcal{X}$) by definition means that there is an approximate solution $\boldsymbol{x}^*$ to (P) that lies in the embedded d-dimensional subspace. One needs to perform an additional global search over the subspace to locate $\boldsymbol{x}^*$. Therefore, for an entirely fair comparison between the two approaches, this additional computational complexity should be taken into account.

5.4 Summary

This chapter has investigated the reduced problem (RP$\mathcal{X}$) applied to the constrained problem (P) for Lipschitz continuous functions that may not have an effective subspace. By looking at the problem through the prism of conic geometry, we developed a new type of analysis to bound the probability of ϵ-success of (RP$\mathcal{X}$). The bounds were expressed in terms of the so-called conic intrinsic volumes of circular cones which have exact formulae and thus are quantifiable. Using these formulae, we analysed the asymptotic behaviour of the bounds for large D. The analysis suggests that the success rate of (RP$\mathcal{X}$) (as expected) decreases exponentially with growing D. Confirming our intuition, the analysis also shows that (RP$\mathcal{X}$) has a high success rate for larger d and smaller distances between the location where subspaces are embedded (i.e. point $\boldsymbol{p}$) and the location of a global minimizer $\boldsymbol{x}^*$. This latter property of (RP$\mathcal{X}$) for general Lipschitz continuous functions is reminiscent of the dependence of success rates of (RP) and (RP$\mathcal{X}$) for functions with low effective dimensionality on the distance between $\boldsymbol{p}_\top$ and $\boldsymbol{x}^*_\top$. Furthermore, to understand the relative performance of (RP$\mathcal{X}$), we compared it with a uniform sampling technique. We looked at lower bounds for the probability of ϵ-success of those two techniques and found that the lower bound $\tau(r_{\boldsymbol{p}})$ for (RP$\mathcal{X}$) is greater than the lower bound τ_{us} for uniform sampling if the distance $\|\boldsymbol{x}^* - \boldsymbol{p}\|$ is smaller than $0.48\sqrt{D}$ in the asymptotic regime ($D \to \infty$). In the asymptotic analysis, the embedding subspace d was kept fixed. The analysis showed that in this regime d has no significant effect on the relative performance of (RP$\mathcal{X}$). Future research may involve comparison of the performances of (RP$\mathcal{X}$) and uniform sampling in different asymptotic settings, for example, when $d = \beta D$ for some fixed constant β.

Our derivations are conceptual in nature and are intended to introduce new tools into the field of global optimization and open up new avenues of research. For example, we believe that it is possible to use the analysis developed in this chapter to obtain lower bounds — that are independent of D — for the probability of ϵ-success of (RP) for functions with low effective dimensionality in the case $d < d_e$. We hope that this analysis will be exploited algorithmically and will allow lifting the restriction of needing to know d_e for random embeddings algorithms for functions with low effective dimensionality.

In the next chapter, we introduce a new algorithmic framework to solve (P) based on the reduced problem (RP$\mathcal{X}$). The framework can be applied to functions with low effective dimensionality as well as general objectives.

Chapter 6

The X-REGO algorithmic framework

This chapter introduces a new generic and flexible algorithmic framework X-REGO for solving (P) for objective functions with or without low effective dimensionality. In Section 6.1, we provide a detailed description of the algorithm and, in Section 6.2, we show that X-REGO is globally convergent under certain mild assumptions and derive rates for functions with coordinate-aligned low-effective subspaces and generic Lipschitz continuous functions. Later, we conduct numerical experiments to test the effectiveness of X-REGO for solving (P) on the set of functions with low effective dimensionality.

6.1 Description of the X-REGO algorithm

In the case of random embeddings for unconstrained global optimization (UP), the success probability of the reduced problem (RP) is independent of the ambient dimension. However, for the constrained problem (P), the analysis in Chapters 4 and 5 indicates that the probability of success of the reduced problem (RP$\mathcal{X}$) decreases with D regardless of the presence of low effective dimensionality. It is thus imperative in any algorithm that uses feasible random embeddings in order to solve (P) to allow multiple such subspaces to be explored, and it is practically important to find out what are efficient and theoretically-sound ways to choose these subspaces iteratively. This is the aim of our generic and flexible algorithmic framework, X-REGO (Algorithm 2). Furthermore, as an additional level of generality and practicality, we allow the reduced, random subproblem to be solved stochastically, so that a sufficiently accurate global solution of this problem is only guaranteed with a certain probability. This covers the obvious case when a (convergent) stochastic global optimization algorithm would be employed to solve the reduced subproblem, but also when a deterministic global solver is used but may sometimes fail to find the required solution due to a limited computational budget, processor failure and so on.

Algorithm 2 $\mathcal{X}$-Random Embeddings for Global Optimization (X-REGO) applied to (P)

```
Initialize $d$ and $\boldsymbol{p}^0 \in \mathcal{X}$
for $k \geq 1$ until termination do
    Draw $\tilde{\boldsymbol{A}}^k$, a realization of the $D \times d$ Gaussian matrix $\boldsymbol{A}$
    Calculate $\tilde{\boldsymbol{y}}^k$ by solving approximately and possibly, probabilistically,
        $$\begin{aligned} \tilde{f}_{min}^k = \min_{\boldsymbol{y} \in \mathbb{R}^d} \ & f(\tilde{\boldsymbol{A}}^k \boldsymbol{y} + \tilde{\boldsymbol{p}}^{k-1}) \\ \text{subject to } & \tilde{\boldsymbol{A}}^k \boldsymbol{y} + \tilde{\boldsymbol{p}}^{k-1} \in \mathcal{X} \end{aligned} \qquad (\widetilde{\mathrm{RP}\mathcal{X}^k})$$
    Let
        $$\tilde{\boldsymbol{x}}^k := \tilde{\boldsymbol{A}}^k \tilde{\boldsymbol{y}}^k + \tilde{\boldsymbol{p}}^{k-1} \qquad (6.1)$$
    Choose (deterministically or randomly) $\tilde{\boldsymbol{p}}^k \in \mathcal{X}$
end for
```

Remark 6.1. Here, we emphasize again that X-REGO can be applied to a general, continuous objective f in (P). If X-REGO is applied to f with low effective dimensionality, the algorithm should enjoy faster convergence. For example, for $\mathcal{T}$ aligned with coordinate axes, the below analysis (Theorem 6.12) shows that the number of iterations required for X-REGO to converge to the optimal set (in probability) is dependent on D algebraically.

In X-REGO, for $k \geq 1$, the kth embedding is determined by a realization $\tilde{\boldsymbol{A}}^k = \boldsymbol{A}^k(\boldsymbol{\omega}^k)$ of the random Gaussian matrix $\boldsymbol{A}^k$, and it is drawn at the point $\tilde{\boldsymbol{p}}^{k-1} = \boldsymbol{p}^{k-1}(\boldsymbol{\omega}^{k-1}) \in \mathcal{X}$, a realization of the random variable $\boldsymbol{p}^{k-1}$ (which, without loss of generality, includes the case of deterministic choices by writing $\boldsymbol{p}^{k-1}$ as a random variable with support equal to a singleton).

X-REGO can be seen as a stochastic process, so that in addition to $\tilde{\boldsymbol{p}}^k$ and $\tilde{\boldsymbol{A}}^k$, each algorithm realization provides sequences $\tilde{\boldsymbol{x}}^k = \boldsymbol{x}^k(\boldsymbol{\omega}^k)$, $\tilde{\boldsymbol{y}}^k = \boldsymbol{y}^k(\boldsymbol{\omega}^k)$ and $\tilde{f}_{min}^k = f_{min}^k(\boldsymbol{\omega}^k)$, for $k \geq 1$, that are realizations of the random variables $\boldsymbol{x}^k$, $\boldsymbol{y}^k$ and f_{min}^k, respectively. Each iteration of X-REGO solves – approximately and possibly, with a certain probability – a realization $(\widetilde{\mathrm{RP}\mathcal{X}^k})$ of the random problem

$$\begin{aligned} f_{min}^k = \min_{\boldsymbol{y}} \ & f(\boldsymbol{A}^k \boldsymbol{y} + \boldsymbol{p}^{k-1}) \\ \text{subject to } & \boldsymbol{A}^k \boldsymbol{y} + \boldsymbol{p}^{k-1} \in \mathcal{X}. \end{aligned} \qquad (\mathrm{RP}\mathcal{X}^k)$$

To calculate $\tilde{\boldsymbol{y}}^k$, $(\widetilde{\mathrm{RP}\mathcal{X}^k})$ may be solved to some required accuracy using a deterministic global optimization algorithm that is allowed to fail with a certain probability; or employing a stochastic algorithm, so that $\tilde{\boldsymbol{y}}^k$ is only guaranteed to be an approximate global minimizer of $(\widetilde{\mathrm{RP}\mathcal{X}^k})$ (at least) with a certain probability.

Several variants of X-REGO can be obtained by specific choices of the random variable $\boldsymbol{p}^k$. A first possibility consists in simply defining $\boldsymbol{p}^k$ as a random variable with support

$\{\mathbf{0}\}$, so that $\tilde{\boldsymbol{p}}^k = \mathbf{0}$ for all k. It is also possible to preserve the progress achieved so far by defining $\boldsymbol{p}^k = \boldsymbol{x}^k_{opt}$, where

$$\boldsymbol{x}^k_{opt} := \arg\min\{f(\boldsymbol{x}^1), f(\boldsymbol{x}^2), \dots, f(\boldsymbol{x}^k)\}, \tag{6.2}$$

the random variable corresponding to the best point found over the k first embeddings. We compare numerically several choices of $\boldsymbol{p}$ on functions with low effective dimensionality in Section 6.3.

The termination in Line 2 could be set to a given maximum number of embeddings, or could check that no significant progress in decreasing the objective function has been achieved over the last few embeddings, compared to the value $f(\tilde{\boldsymbol{x}}^k_{opt})$. For generality, we leave it unspecified here.

6.2 Global convergence of the X-REGO algorithm to the set of global ϵ-minimizers

For a(ny) given tolerance $\epsilon > 0$, let G_ϵ be the set of approximate global minimizers of (P) defined in (4.73). We show that $\boldsymbol{x}^k_{opt}$ in (6.2) converges to G_ϵ almost surely as $k \to \infty$ (see Theorem 6.9).

Intuitively, our proof relies on the fact that any vector $\tilde{\boldsymbol{x}}^k$ defined in (6.1) belongs to G_ϵ if the following two conditions hold simultaneously: (a) the reduced problem (RP$\mathcal{X}^k$) is $(\epsilon - \lambda)$-successful in the sense of Definition 4.24[1], namely,

$$f^k_{min} \leq f^* + \epsilon - \lambda; \tag{6.3}$$

(b) the reduced problem $(\widetilde{\text{RP}\mathcal{X}^k})$ is solved (by a deterministic/stochastic algorithm) to an accuracy $\lambda \in (0, \epsilon)$ in the objective function value, namely,

$$f(\boldsymbol{A}^k \boldsymbol{y}^k + \boldsymbol{p}^{k-1}) \leq f^k_{min} + \lambda \tag{6.4}$$

holds (at least) with a certain probability. We introduce two additional random variables that capture the conditions in (a) and (b) above,

$$R^k = \mathbb{1}\{(\text{RP}\mathcal{X}^k) \text{ is } (\epsilon - \lambda)\text{-successful in the sense of (6.3)}\}, \tag{6.5}$$

$$S^k = \mathbb{1}\{(\text{RP}\mathcal{X}^k) \text{ is solved to accuracy } \lambda \text{ in the sense of (6.4)}\}, \tag{6.6}$$

where $\mathbb{1}$ is the usual indicator function for an event.

---

[1] The reader may expect us to simply require that (RP$\mathcal{X}^k$) is ϵ-successful. However, in order to ensure convergence of X-REGO to the set of ϵ-minimizers, we need to be slightly more demanding on the success requirements for (RP$\mathcal{X}^k$) so that we allow inexact solutions (up to accuracy λ) of the reduced problem $(\widetilde{\text{RP}\mathcal{X}^k})$.

Let $\mathcal{F}^k = \sigma(\boldsymbol{A}^1, \dots, \boldsymbol{A}^k, \boldsymbol{y}^1, \dots, \boldsymbol{y}^k, \boldsymbol{p}^0, \dots, \boldsymbol{p}^k)$ be the σ-algebra generated by the random variables $\boldsymbol{A}^1, \dots, \boldsymbol{A}^k, \boldsymbol{y}^1, \dots, \boldsymbol{y}^k, \boldsymbol{p}^0, \dots, \boldsymbol{p}^k$ (a mathematical concept that represents the history of the X-REGO algorithm as well as its randomness until the kth embedding)[2], with $\mathcal{F}^0 = \sigma(\boldsymbol{p}^0)$. We also construct an 'intermediate' σ-algebra, namely,

$$\mathcal{F}^{k-1/2} = \sigma(\boldsymbol{A}^1, \dots, \boldsymbol{A}^{k-1}, \boldsymbol{A}^k, \boldsymbol{y}^1, \dots, \boldsymbol{y}^{k-1}, \boldsymbol{p}^0, \dots, \boldsymbol{p}^{k-1}),$$

with $\mathcal{F}^{1/2} = \sigma(\boldsymbol{p}^0, \boldsymbol{A}^1)$. Note that $\boldsymbol{x}^k$, R^k and S^k are $\mathcal{F}^k$-measurable[3], and R^k is also $\mathcal{F}^{k-1/2}$-measurable; thus they are well-defined random variables.

Remark 6.2. The random variables $\boldsymbol{A}^1, \dots, \boldsymbol{A}^k$, $\boldsymbol{y}^1, \dots, \boldsymbol{y}^k$, $\boldsymbol{x}^1, \dots, \boldsymbol{x}^k$, $\boldsymbol{p}^0, \boldsymbol{p}^1, \dots, \boldsymbol{p}^k$, R^1, $\dots$, R^k, $S^1, \dots, S^k$ are $\mathcal{F}^k$-measurable since $\mathcal{F}^0 \subseteq \mathcal{F}^1 \subseteq \cdots \subseteq \mathcal{F}^k$. Also, $\boldsymbol{A}^1, \dots, \boldsymbol{A}^k$, $\boldsymbol{y}^1, \dots, \boldsymbol{y}^{k-1}$, $\boldsymbol{x}^1, \dots, \boldsymbol{x}^{k-1}$, $\boldsymbol{p}^0, \boldsymbol{p}^1, \dots, \boldsymbol{p}^{k-1}$, $R^1, \dots, R^k$, $S^1, \dots, S^{k-1}$ are $\mathcal{F}^{k-1/2}$-measurable since $\mathcal{F}^0 \subseteq \mathcal{F}^{1/2} \subseteq \mathcal{F}^1 \subseteq \cdots \subseteq \mathcal{F}^{k-1} \subseteq \mathcal{F}^{k-1/2}$.

A weak assumption is given next, that is satisfied by reasonable techniques for the subproblems; namely, the reduced problem (RP$\mathcal{X}^k$) needs to be solved to required accuracy with some positive probability.

Assumption 6.3. There exists $\rho \in (0, 1]$ such that, for all $k \geq 1$,[4]

$$\mathbb{P}[S^k = 1 | \mathcal{F}^{k-1/2}] = \mathbb{E}[S^k | \mathcal{F}^{k-1/2}] \geq \rho,$$

i.e., with (conditional) probability at least $\rho > 0$, the solution $\boldsymbol{y}^k$ of (RP$\mathcal{X}^k$) satisfies (6.4).

Remark 6.4. If a deterministic (global optimization) algorithm is used to solve $(\widetilde{\text{RP}\mathcal{X}^k})$, then S^k is always $\mathcal{F}^{k-1/2}$-measurable and Assumption 6.3 is equivalent to $S^k \geq \rho$. Since S^k is an indicator function, this further implies that $S^k \equiv 1$.

For X-REGO to converge, it is also necessary that the drawn subspaces intersect the set of global minimizers G_ϵ with a positive probability. This property holds for Lipschitz continuous functions with low effective dimensionality (as proven in Theorem 4.25) and without (see Theorem 5.16). For generality, we state this property as an assumption.

Assumption 6.5. There exists $\tau \in (0, 1]$ such that, for all $k \geq 1$,

$$\mathbb{P}[R^k = 1 | \mathcal{F}^{k-1}] = \mathbb{E}[R^k | \mathcal{F}^{k-1}] \geq \tau, \tag{6.7}$$

i.e., with (conditional) probability at least $\tau > 0$, (RP$\mathcal{X}^k$) is $(\epsilon - \lambda)$-successful.

---

[2] A similar setup for random iterates of probabilistic models can be found in [12, 30].

[3] It would be possible to restrict the definition of the σ-algebra $\mathcal{F}^k$ so that it contains strictly the randomness of the embeddings $\boldsymbol{A}^i$ and $\boldsymbol{p}^i$ for $i \leq k$; then we would need to assume that $\boldsymbol{y}^k$ is $\mathcal{F}^k$-measurable, which would imply that R^k, S^k and $\boldsymbol{x}^k$ are also $\mathcal{F}^k$-measurable. Similar comments apply to the definition of $\mathcal{F}^{k-1/2}$.

[4] The equality in the displayed equation follows from $\mathbb{E}[S^k|\mathcal{F}^{k-1}] = 1 \cdot \mathbb{P}[S^k = 1|\mathcal{F}^{k-1}] + 0 \cdot \mathbb{P}[S^k = 0|\mathcal{F}^{k-1}]$.

6.2.1 A generic proof of global convergence of X-REGO

We will need the following three lemmas to assist the proof of convergence.

Lemma 6.6. *If Assumptions 6.3 holds, then*

$$\mathbb{E}[R^k S^k | \mathcal{F}^{k-1/2}] \geq \rho R^k, \quad \text{for} \quad k \geq 1. \tag{6.8}$$

Proof. Assumption 6.3 implies

$$\mathbb{E}[R^k S^k | \mathcal{F}^{k-1/2}] = R^k \mathbb{E}[S^k | \mathcal{F}^{k-1/2}] \geq \rho R^k,$$

where the equality follows from the fact that R^k is $\mathcal{F}^{k-1/2}$-measurable and, thus, can be pulled out of the expectation (see [46, Theorem 4.1.14]). □

A useful property is given next.

Lemma 6.7. *Let Assumptions LipC, 6.3 and 6.5 hold. Then, for $K \geq 1$, we have*

$$\mathbb{P}\Big[\bigcup_{k=1}^{K} \{ \{R^k = 1\} \cap \{S^k = 1\} \} \Big] \geq 1 - (1 - \tau\rho)^K.$$

Proof. We define an auxiliary random variable, $J^K := \mathbb{1}\left(\bigcup_{k=1}^{K} \{ \{R^k = 1\} \cap \{S^k = 1\} \} \right)$. Note that $J^K = 1 - \prod_{k=1}^{K} (1 - R^k S^k)$. We have

$$\begin{aligned}
\mathbb{P}\Big[\bigcup_{k=1}^{K} \{ \{R^k = 1\} \cap \{S^k = 1\} \} \Big] &= \mathbb{E}[J^K] = 1 - \mathbb{E}\Big[\prod_{k=1}^{K} (1 - R^k S^k) \Big] \\
&\overset{(*)}{=} 1 - \mathbb{E}\Big[\mathbb{E}\Big[\prod_{k=1}^{K} (1 - R^k S^k) \Big| \mathcal{F}^{K-1/2} \Big] \Big] \\
&\overset{(\circ)}{=} 1 - \mathbb{E}\Big[\prod_{k=1}^{K-1} (1 - R^k S^k) \cdot \mathbb{E}\big[1 - R^K S^K | \mathcal{F}^{K-1/2} \big] \Big] \\
&\geq 1 - \mathbb{E}\Big[(1 - \rho R^K) \cdot \prod_{k=1}^{K-1} (1 - R^k S^k) \Big] \\
&\overset{(*)}{=} 1 - \mathbb{E}\Big[\mathbb{E}\Big[(1 - \rho R^K) \cdot \prod_{k=1}^{K-1} (1 - R^k S^k) \Big| \mathcal{F}^{K-1} \Big] \Big] \\
&\overset{(\circ)}{=} 1 - \mathbb{E}\Big[\prod_{k=1}^{K-1} (1 - R^k S^k) \cdot \mathbb{E}\big[1 - \rho R^K | \mathcal{F}^{K-1} \big] \Big] \\
&\geq 1 - (1 - \tau\rho) \cdot \mathbb{E}\Big[\prod_{k=1}^{K-1} (1 - R^k S^k) \Big],
\end{aligned}$$

where

- $(*)$ follow from the tower property of conditional expectation (see (4.1.5) in [46]),
- $(\circ)$ is due to the fact that $R^1, \ldots, R^{K-1}$ and $S^1, \ldots, S^{K-1}$ are $\mathcal{F}^{K-1/2}$- and $\mathcal{F}^{K-1}$-measurable (see Theorem 4.1.14 in [46]),
- the inequalities follow from (6.8) and (6.7), respectively.

We repeatedly expand the expectation of the product for $K-1$, ..., 1, in exactly the same manner as above, to obtain the desired result. □

In the next lemma, we show that if (RP$\mathcal{X}^k$) is $(\epsilon-\lambda)$-successful and is solved to accuracy λ in objective value, then the solution $\boldsymbol{x}^k$ must be inside G_ϵ; thus proving our intuitive statements (a) and (b) at the start of Section 6.2.

Lemma 6.8. *Suppose Assumptions LipC, 6.3 and 6.5 hold. Then,*

$$\{R^k = 1\} \cap \{S^k = 1\} \subseteq \{\boldsymbol{x}^k \in G_\epsilon\}.$$

Proof. By Definition 4.24, if (RP$\mathcal{X}^k$) is $(\epsilon - \lambda)$-successful, then there exists $\boldsymbol{y}^k_{int} \in \mathbb{R}^d$ such that $\boldsymbol{A}^k \boldsymbol{y}^k_{int} + \boldsymbol{p}^{k-1} \in \mathcal{X}$ and

$$f(\boldsymbol{A}^k \boldsymbol{y}^k_{int} + \boldsymbol{p}^{k-1}) \leq f^* + \epsilon - \lambda. \tag{6.9}$$

Since $\boldsymbol{y}^k_{int}$ is in the feasible set of (RP$\mathcal{X}^k$) and f^k_{min} is the global minimum of (RP$\mathcal{X}^k$), we have

$$f^k_{min} \leq f(\boldsymbol{A}^k \boldsymbol{y}^k_{int} + \boldsymbol{p}^{k-1}). \tag{6.10}$$

Then, for $\boldsymbol{x}^k$, (6.4) gives the first inequality below,

$$f(\boldsymbol{x}^k) \leq f^k_{min} + \lambda \leq f(\boldsymbol{A}^k \boldsymbol{y}^k_{int} + \boldsymbol{p}^{k-1}) + \lambda \leq f^* + \epsilon,$$

where the second and third inequalities follow from (6.10) and (6.9), respectively. This shows that $\boldsymbol{x}^k \in G_\epsilon$. □

The next theorem is the culmination of this chapter's analysis. With the above lemmas at hand, we are now ready to prove that X-REGO is globally convergent.

Theorem 6.9 (Global convergence). *Suppose Assumptions LipC, 6.3 and 6.5 hold. Then,*

$$\lim_{k\to\infty} \mathbb{P}[\boldsymbol{x}^k_{opt} \in G_\epsilon] = \lim_{k\to\infty} \mathbb{P}[f(\boldsymbol{x}^k_{opt}) \leq f^* + \epsilon] = 1 \tag{6.11}$$

where $\boldsymbol{x}^k_{opt}$ and G_ϵ are defined in (6.2) and (4.73), respectively.

Furthermore, for any $\xi \in (0,1)$,

$$\mathbb{P}[\boldsymbol{x}^k_{opt} \in G_\epsilon] = \mathbb{P}[f(\boldsymbol{x}^k_{opt}) \leq f^* + \epsilon] \geq \xi \textit{ for all } k \geq K_\xi, \tag{6.12}$$

where $K_\xi := \left\lceil \dfrac{|\log(1-\xi)|}{\tau\rho} \right\rceil$.

Proof. Lemma 6.8 and the definition of $\boldsymbol{x}_{opt}^k$ in (6.2) provide

$$\{R^k = 1\} \cap \{S^k = 1\} \subseteq \{\boldsymbol{x}^k \in G_\epsilon\} \subseteq \{\boldsymbol{x}_{opt}^k \in G_\epsilon\}$$

for $k = 1, 2, \dots, K$ and for any integer $K \geq 1$. Hence,

$$\bigcup_{k=1}^{K} \{R^k = 1\} \cap \{S^k = 1\} \subseteq \bigcup_{k=1}^{K} \{\boldsymbol{x}_{opt}^k \in G_\epsilon\}. \tag{6.13}$$

Note that the sequence $\{f(\boldsymbol{x}_{opt}^1), f(\boldsymbol{x}_{opt}^2), \dots, f(\boldsymbol{x}_{opt}^K)\}$ is monotonically decreasing. Therefore, if $\boldsymbol{x}_{opt}^k \in G_\epsilon$ for some $k \leq K$ then $\boldsymbol{x}_{opt}^i \in G_\epsilon$ for all $i = k, \dots, K$; and so the sequence $(\{\boldsymbol{x}_{opt}^k \in G_\epsilon\})_{k=1}^K$ is an increasing sequence of events. Hence,

$$\bigcup_{k=1}^{K} \{\boldsymbol{x}_{opt}^k \in G_\epsilon\} = \{\boldsymbol{x}_{opt}^K \in G_\epsilon\}. \tag{6.14}$$

From (6.14) and (6.13), we have for all $K \geq 1$,

$$\mathbb{P}[\{\boldsymbol{x}_{opt}^K \in G_\epsilon\}] \geq \mathbb{P}\Big[\bigcup_{k=1}^{K} \{R^k = 1\} \cap \{S^k = 1\}\Big] \geq 1 - (1 - \tau\rho)^K, \tag{6.15}$$

where the second inequality follows from Lemma 6.7. Finally, passing to the limit with K in (6.15), we deduce $1 \geq \lim_{K\to\infty} \mathbb{P}[\{\boldsymbol{x}_{opt}^K \in G_\epsilon\}] \geq \lim_{K\to\infty} \left[1 - (1 - \tau\rho)^K\right] = 1$, as required.

Note that if

$$1 - (1 - \tau\rho)^k \geq \xi \tag{6.16}$$

then (6.15) implies $\mathbb{P}[\boldsymbol{x}_{opt}^k \in G_\epsilon] \geq \xi$. Since (6.16) is equivalent to $k \geq \dfrac{\log(1-\xi)}{\log(1-\tau\rho)}$, (6.16) holds for all $k \geq K_\xi$ since $K_\xi \geq \dfrac{\log(1-\xi)}{\log(1-\tau\rho)}$. □

Remark 6.10. If f is a convex function (and known a priori to be so), then clearly, a local (deterministic or stochastic) optimization algorithm may be used to solve $(\widetilde{\mathrm{RP}\mathcal{X}^k})$ and achieve (6.4). Apart from this important speed-up and simplification, it is difficult to exploit this additional special structure of f in our analysis, in order to improve the success bounds and convergence.

In the following two sections, we draw implications of Theorem 6.9 for functions with low effective dimensionality and general Lipschitz continuous functions. In particular, we replace Assumption 6.5 with Theorem 4.25 and Theorem 5.16, and quantify rates of convergence based on estimations of respective τ in Sections 4.2.2 and 5.3.1.

6.2.2 Global convergence of X-REGO for functions with low effective dimensionality

We first present a corollary of Theorem 6.9 for functions with low effective dimensionality. Note that, for Lipschitz continuous functions with low effective dimensionality, Theorem 4.25 implies (6.7) if we additionally assume that there is a set G^* that is non-degenerate.

Corollary 6.11. *Suppose that Assumptions LowED, LipC and 6.3 hold. Suppose also that there is a set G^* defined in Definition 2.20 that is non-degenerate according to Definition 4.10. Furthermore, assume that $d \geq d_e$ for all drawn matrices $\boldsymbol{A}^1, \boldsymbol{A}^2, \dots$. Then,* (6.11) *holds.*

Proof. We just need to show that with the given assumptions, (6.7) in Assumption 6.5 is satisfied. Then, the results follow from Theorem 6.9.

Recall that the support of the random variable $\boldsymbol{p}^k$ is contained in $\mathcal{X}$. For each embedding, we apply Theorem 4.25 (setting $\boldsymbol{p} = \tilde{\boldsymbol{p}}^{k-1}$ and replacing ϵ by $\epsilon - \lambda$) to deduce that there exists $\tau \in (0, 1]$ such that $\mathbb{P}[R^k = 1|\mathcal{F}^{k-1}] \geq \tau$, for $k \geq 1$. Equivalently, in terms of conditional expectation, we can write $\mathbb{E}[R^k|\mathcal{F}^{k-1}] = 1 \cdot \mathbb{P}[R^k = 1|\mathcal{F}^{k-1}] + 0 \cdot \mathbb{P}[R^k = 0|\mathcal{F}^{k-1}] \geq \tau$. This shows that (6.7) in Assumption 6.5 holds. □

Coordinate-aligned effective subspace. Using the estimates for τ in Theorem 4.13, we can estimate precisely the rate of convergence of X-REGO as a function of problem dimension, assuming that $\mathcal{T}$ is aligned with coordinate axes.

Theorem 6.12. *Suppose Assumption LowED holds with $\boldsymbol{U} = [\boldsymbol{I}_{d_e}\ \boldsymbol{0}]^T$ and $\boldsymbol{V} = [\boldsymbol{0}\ \boldsymbol{I}_{D-d_e}]^T$, as well as Assumption 6.3. Let $\xi \in (0, 1)$, and d_e and d be fixed with $d \geq d_e$. Then,* (6.12) *holds with*

$$K_\xi = \frac{|\log(1-\xi)|}{\rho} O\left(\frac{2^{D-d_e} \cdot (D - d_e + 1)^{d_e}}{\log(D - d_e + 1)^{\frac{d-1}{2}}}\right) \text{ as } D \to \infty. \tag{6.17}$$

If $\boldsymbol{p}^k = \boldsymbol{0}$ for $k \geq 0$, then (6.12) *holds with*

$$K_\xi = \frac{|\log(1-\xi)|}{\rho} O\left(\frac{(D - d_e + 1)^{d_e}}{\log(D - d_e + 1)^{\frac{d-1}{2}}}\right) \text{ as } D \to \infty. \tag{6.18}$$

Proof. Firstly, note our remark regarding assumptions below. The result follows from Theorem 6.9, (4.25) and (4.27). □

Remark 6.13. Assumption LipC and the assumption that there is a non-degenerate set G^* were required to prove Theorem 4.25 and, consequently, (6.7). If the effective subspace is aligned with coordinate axes, we no longer need these two assumptions to prove (6.7).

In this case, (6.7) follows from Theorem 4.13, together with the fact that (RP$\mathcal{X}^k$) being successful implies (RP$\mathcal{X}^k$) is ϵ-successful for any $\epsilon \geq 0$.

6.2.3 Global convergence of X-REGO for general objectives

We present a similar corollary for general Lipschitz continuous functions. We replace Assumption 6.5 with Theorem 5.16 and Assumption 5.12, which says that there exists a minimizer $\boldsymbol{x}^*$ that is sufficiently away from the boundary of $\mathcal{X}$.

Corollary 6.14. *Suppose that Assumptions LipC, 5.12 and 6.3 hold. Then,* (6.11) *holds.*

Proof. The proof is identical to the proof of Corollary 6.11 except that we now use Theorem 5.16 to imply (6.7) in Assumption 6.5. ☐

Similar to Theorem 6.12, we estimate the rate of convergence of X-REGO for Lipschitz continuous functions using the estimates for τ provided in Corollary 5.20.

Theorem 6.15. *Suppose that Assumptions LipC, 5.12 and 6.3 hold. Then,* (6.12) *holds with*

$$K_\xi = \frac{|\log(1-\xi)|}{\rho} O\left(D^{\frac{D}{2}-d+1} \left(\frac{2L}{\epsilon}\right)^{D-d} \right) \textit{ as } D \to \infty. \tag{6.19}$$

Proof. The result follows from Theorem 6.9 and (5.41). ☐

6.3 Numerical experiments

Let us now assess performance of X-REGO (Algorithm 2) numerically. We shall conduct the experiments on the set of functions with low effective dimensionality.

6.3.1 Setup

Algorithms. We test different variants of Algorithm 2 against the *no-embedding* framework, in which (P) is solved directly without using random embeddings and with no explicit exploitation of its special structure. Each variant of X-REGO corresponds to a specific choice of $\boldsymbol{p}^k$, $k \geq 0$:

- Adaptive X-REGO (A-REGO). In X-REGO, the point $\boldsymbol{p}^k$ is chosen as the best point found up to the kth embedding: if $f(\boldsymbol{A}^k\boldsymbol{y}^k + \boldsymbol{p}^{k-1}) < f(\boldsymbol{p}^{k-1})$ then $\boldsymbol{p}^k := \boldsymbol{A}^k\boldsymbol{y}^k + \boldsymbol{p}^{k-1}$, otherwise, $\boldsymbol{p}^k := \boldsymbol{p}^{k-1}$.
- Local Adaptive X-REGO (LA-REGO). In X-REGO, we solve ($\widetilde{\text{RP}\mathcal{X}^k}$) using a local solver (instead of a global one as in A-REGO). Then, if $f(\boldsymbol{p}^{k-1}) - f(\boldsymbol{A}^k\boldsymbol{y}^k + \boldsymbol{p}^{k-1}) > \gamma$ for some small γ (here, $\gamma = 10^{-5}$), we let $\boldsymbol{p}^k := \boldsymbol{A}^k\boldsymbol{y}^k + \boldsymbol{p}^{k-1}$, otherwise, $\boldsymbol{p}^k$ is chosen uniformly at random in $\mathcal{X}$.

- Nonadaptive X-REGO (N-REGO). In X-REGO, all the random subspaces are drawn at the origin: $\boldsymbol{p}^k := \boldsymbol{0}$ for all k.

- Local Nonadaptive X-REGO (LN-REGO). In X-REGO, the low-dimensional problem ($\widetilde{\mathrm{RP}\mathcal{X}^k}$) is solved using a local solver, and the point $\boldsymbol{p}^k$ is chosen uniformly at random in $\mathcal{X}$ for all k.

Solvers. We test the aforementioned X-REGO variants using three solvers for solving the reduced problem ($\widetilde{\mathrm{RP}\mathcal{X}^k}$) (or the original problem (P) in the no-embedding case), namely, DIRECT ([52, 58, 78]), BARON ([139, 148]) and KNITRO ([28]).

We test A-REGO and N-REGO using DIRECT, BARON and multi-start KNITRO (referred to here as mKNITRO) and test LA-REGO and LN-REGO using only KNITRO with no multi-start feature.

Test set. To assess Algorithm 2, we use the same test set we used in Section 3.4 to assess Algorithm 1. Refer to Section 3.3 for the description of how we constructed our test set of functions with low effective dimensionality.

Experimental setup. For each version of X-REGO and its paired solvers, we solve the entire test set 5 times to estimate the average performance of the algorithms. Let f be a function from the test set with the global minimum f^*. When applying any version of X-REGO to minimize f, we terminate either after $K = 100$ embeddings, or earlier, as soon as[5]

$$f(\widetilde{\boldsymbol{A}}^k \tilde{\boldsymbol{y}}^k + \tilde{\boldsymbol{p}}^{k-1}) - f^* \leq \epsilon = 10^{-3}. \tag{6.20}$$

We then record the computational cost, which we measure in terms of either function evaluations or CPU time in seconds. To compare with 'no-embedding', we solve the full-dimensional problem (P) directly with DIRECT, BARON and mKNITRO with no use of random embeddings. The budget and termination criteria used for each solver to solve ($\widetilde{\mathrm{RP}\mathcal{X}^k}$) within X-REGO or to solve (P) in the 'no-embedding' framework are outlined in Table 3.1.

Remark 6.16. The experiments are done not to compare solvers but to contrast 'no-embedding' with the X-REGO variants. All the experiments were run in MATLAB on the 16 cores (2×8 Intel with hyper-threading) Linux machines with 256GB RAM and 3300 MHz speed.

---

[5] We acknowledge that the use of the true global minimum f^*, or a sufficiently close lower bound, in our numerical testing is not practical. But we note that our aim here is to test both 'no-embedding' and X-REGO in similar, even if idealized, settings.

Table 6.1: The table outlines the experimental setup for the solvers, used both in the 'no-embedding' algorithm and for solving the low-dimensional problem $(\widetilde{\mathrm{RP}\mathcal{X}^k})$ (as usually, f denotes the D-dimensional function to minimize, f^* is its global minimum, and ϵ in (6.20) is set to 10^{-3}). At each internal iteration, DIRECT stores f^*_{Di} — the minimal value of f found so far, while BARON stores f^U_{Ba} and f^L_{Ba} — the smallest upper bound and largest lower bound found found so far. Note that, for BARON, $f^U_{Ba} = f(\boldsymbol{x}^k)$ in $(\widetilde{\mathrm{RP}\mathcal{X}^k})$.

	DIRECT	BARON	mKNITRO	KNITRO
Measure of computational cost	function evaluations	CPU seconds	function evaluations	function evaluations
Max. budget to solve $(\widetilde{\mathrm{RP}\mathcal{X}^k})$	3000 function evaluations	5 CPU seconds	5 starting points	1 starting point
Max. budget to solve (P)	60000 function evaluations	1000 CPU seconds	100 starting points	Not applicable
Termination for $(\widetilde{\mathrm{RP}\mathcal{X}^k})$	Terminate either on budget or if $f^*_{Di} \leq f^* + \epsilon$	Terminate either on budget or if f^U_{Ba} and f^L_{Ba} satisfy $f^U_{Ba} \leq f^L_{Ba} + \epsilon$	Default options (unless overwritten by additional options)	Default options (unless overwritten by additional options)
Termination for (P)	Same as above	Terminate either on budget or if f^U_{Ba} satisfies $f^U_{Ba} \leq f^* + \epsilon$	Same as above	Not applicable
Additional options for $(\widetilde{\mathrm{RP}\mathcal{X}^k})$	`testflag`=1 `maxits`=Inf `globalmin`=f^*	Default options	`ms_enable`=1 `fstopval`=$f^* + \epsilon$	`fstopval`=$f^* + \epsilon$
Additional options for (P)	Same as above	Same as above	Same as above	Not applicable

We present the main numerical results using Dolan and Moré's performance profile [44] — a popular framework to compare performances of optimization algorithms applied to a given test set. For a given algorithm $\mathcal{A}$, and for each function f in the test set $\mathcal{S}$, we define

$$N_f(\mathcal{A}) := \text{min. \# of fun. evals (or CPU sec.) required to achieve } f(\boldsymbol{x}^k) \leq f^* + \epsilon.$$

If $\mathcal{A}$ fails to successfully converge to the global minimum of f within the maximum computational budget, we set $N_f(\mathcal{A}) = \infty$.

Remark 6.17. Note that, in our experiments, an algorithm $\mathcal{A}$ corresponds to a pair (optimization solver, algorithmic framework) such as (BARON, no embedding) or (KNITRO, N-REGO).

We further define

$$N^*_f := \min_{\mathcal{A}} N_f(\mathcal{A}),$$

as the minimal computational cost by any algorithm required to optimize f. We normalize all the computational costs by the algorithms to solve f by N^*_f and, for each $\mathcal{A}$, we plot

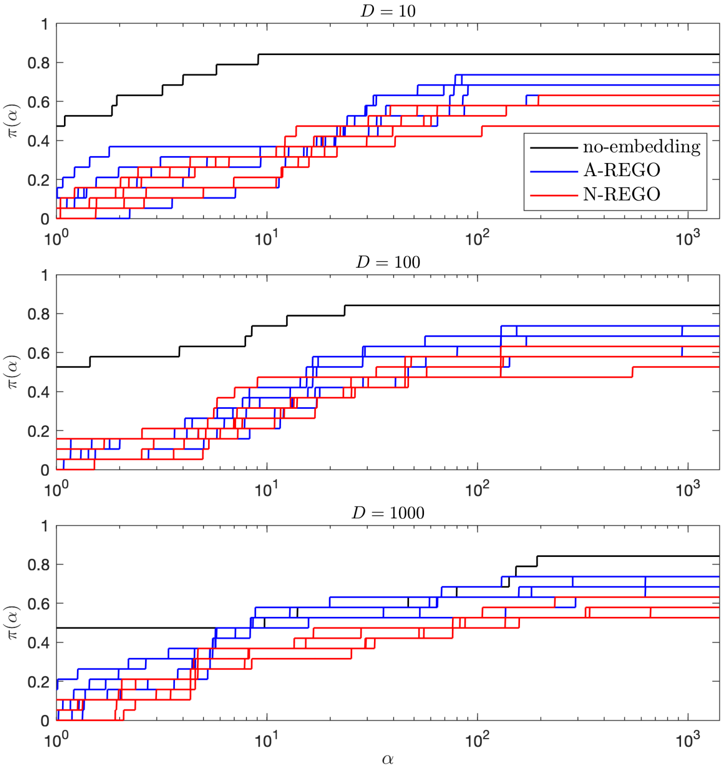


Figure 6.1: Comparison between the X-REGO algorithms and 'no-embedding' with DIRECT.

a function $\pi_{\mathcal{A}}(\alpha)$ that computes the proportion of f's in the test set $\mathcal{S}$, for which the normalized computational effort spent by $\mathcal{A}$ was less than α. Mathematically speaking,

$$\pi_{\mathcal{A}}(\alpha) := \frac{|\{f : N_f(\mathcal{A}) \leq \alpha N_f^*\}|}{|\mathcal{S}|} \text{ for } \alpha \geq 1,$$

where $|\cdot|$ denotes the cardinality of a set. The algorithm $\mathcal{A}$ is considered to have achieved better performance if it produces higher values for $\pi_{\mathcal{A}}(\alpha)$ for lower values of α, i.e., on figures, the curve of $\pi_{\mathcal{A}}(\alpha)$ is higher and lefter.

Remark 6.18. Each time we solve the test set with an X-REGO algorithm coupled with any optimization solver, we obtain different results due to randomness involved in the X-REGO algorithms. Thus, we consider each of the 5 applications of (X-REGO algorithm, optimization solver) to the test set as a separate $\mathcal{A}$; this results in 5 different curves for the same (X-REGO algorithm, optimization solver) pair.

6.3.2 Numerical results

We make separate comparisons for each optimization solver and for each D.

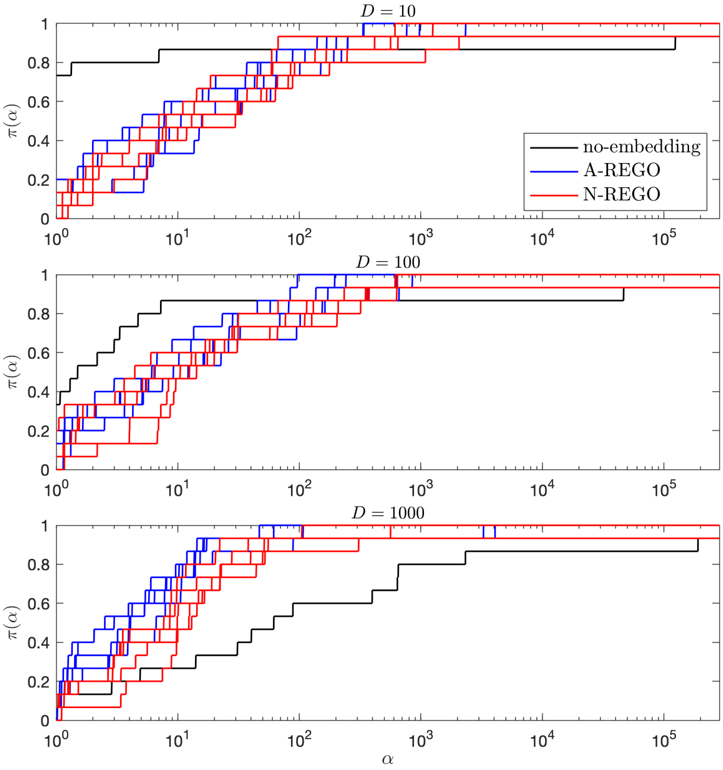


Figure 6.2: Comparison between the X-REGO algorithms and 'no-embedding' with BARON.

DIRECT: Figure 6.1 compares the adaptive and non-adaptive random embedding algorithms (A-REGO and N-REGO) to the no-embedding framework, when using the DIRECT solver for the reduced problem ($\widetilde{\mathrm{RP}\mathcal{X}^k}$) (and for the full-dimensional problem in the case of the no-embedding framework). We find that the no-embedding framework outperforms the two X-REGO variants. We also note that this behaviour is more pronounced when the dimension of the problem (P) is small. In that regime, it is also difficult to determine which version of X-REGO performs the best. When D is large, the no-embedding framework still outperforms the two variants of X-REGO, but among these two, the adaptive one (A-REGO) performs generally better than N-REGO. The median number of function evaluations required by the algorithms, measured over the five repetitions of the experiment, is given in Table 6.2.

BARON: Figure 6.2 compares A-REGO and N-REGO to the no-embedding framework, when using BARON to solve the reduced problem ($\widetilde{\mathrm{RP}\mathcal{X}^k}$). We find that the no-embedding framework is clearly outperformed by the two variants of X-REGO in the large-dimensional setting. Then, it is also clear that the adaptive variant of X-REGO outperforms the non-

Table 6.2: Median number of function evaluations or CPU time spent by each algorithm-solver pair.

	DIRECT (fun. evals)			BARON (CPU sec.)			KNITRO (fun. evals)		
	$D = 10$	$D = 10^2$	$D = 10^3$	$D = 10$	$D = 10^2$	$D = 10^3$	$D = 10$	$D = 10^2$	$D = 10^3$
no-embedding	1261	16933	63795	0.08	0.50	155.20	220	1425	11542
A-REGO	24569	300348	300276	0.63	1.93	15.66	1534	3992	5346
LA-REGO	–	–	–	–	–	–	230	730	4165
N-REGO	63093	300484	300532	0.82	3.00	21.51	1582	3606	8766
LN-REGO	–	–	–	–	–	–	220	763	704

adaptive one. Table 6.2 also indicates that the CPU time used by the different algorithms increases with the dimension of the problem, and that the increase is most rapid for 'no-embedding'.

KNITRO: The comparison between the X-REGO variants, using (m)KNITRO to solve $(\widetilde{\mathrm{RP}\mathcal{X}^k})$, is given in Figure 6.3. Here, we also compare the local variants of X-REGO (namely, LA-REGO and LN-REGO), for which the reduced problem is solved using local KNITRO, with no multi-start feature. We find that the local variants outperform the global ones, and the no-embedding framework when the dimension of the problem is sufficiently large. Figure 6.3 also indicates that the local non-adaptive variant (LN-REGO) outperforms the adaptive one in this high-dimensional setting. This behaviour can also be observed in Table 6.2, which indicates that the median number of function evaluations increases significantly for LA-REGO while for LN-REGO, it actually decreases.

Remark 6.19. We note that two branching type solvers BARON and DIRECT perform branching in a d-dimensional embedded subspace (and not in $\mathbb{R}^D$). When X-REGO begins a new iteration, a random (different) subspace is drawn, and a new branching procedure must be initiated by BARON/DIRECT in this new subspace. X-REGO does not take into account branching done in the previous iterations/subspaces, which could have been used to inform us of the next moves or save computational resources in subsequent iterations. In this regard, it might conceptually and numerically be better to branch in $\mathcal{X}$ box instead. This would, however, require us to look at the random embedding technique anew — when combined with branching type solvers — as its current application only allows branching in $\mathbb{R}^d$.

6.3.3 Conclusions

The numerical experiments indicate that the X-REGO algorithm is mostly beneficial (as expected) for high-dimensional problems (i.e., when D is large). In this setting, the different

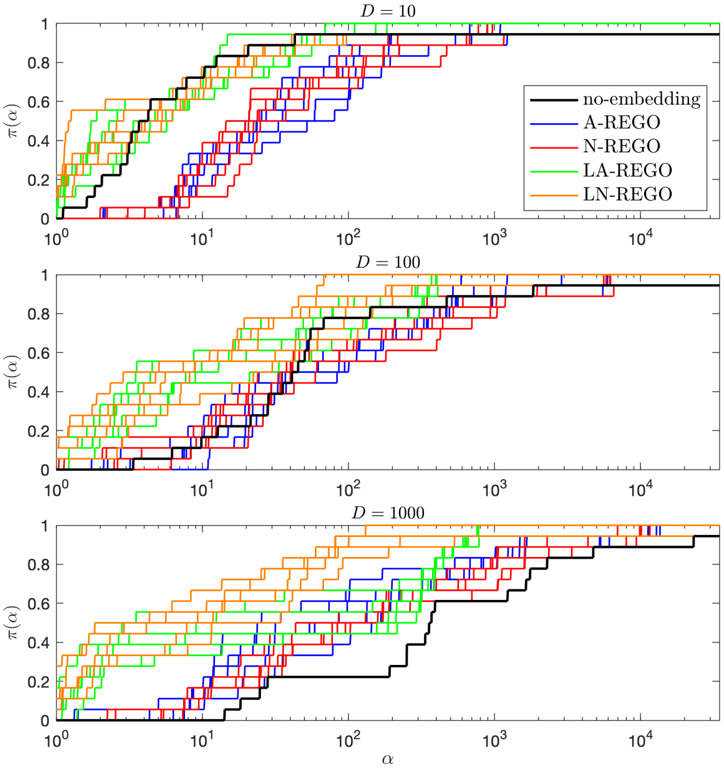


Figure 6.3: Comparison between the X-REGO algorithms and 'no-embedding' with KNITRO.

variants of X-REGO considered here outperform the no-embedding framework when combined with the solvers BARON and mKNITRO. It is less obvious to decide which variant of X-REGO is the best, but it seems that, at least on the problem set considered, the local methods outperform the global ones.

6.4 Summary

This chapter has introduced and analysed a generic and flexible algorithmic framework X-REGO for solving bound-constrained problem (P). X-REGO is based on repetitive application of the reduced problem (RP$\mathcal{X}$), where the parameter $\boldsymbol{p}$ is user-specified. Flexibility in choosing $\boldsymbol{p}$ allows the user to calibrate the level of exploration in $\mathcal{X}$.

The central result of this chapter was the proof of global convergence of X-REGO. Our global convergence result applies to general Lipschitz continuous functions and the presence of low effective dimensionality is not necessary for the global convergence result to hold. We derived different rates of convergence based on estimations of lower bounds for the ϵ-success of (RP$\mathcal{X}$) in Chapters 4 and 5. For functions with low effective dimensionality, we noted that X-REGO should enjoy faster convergence and showed that, for the special case of coordinate-aligned effective subspaces, the rate of convergence is dependent on D algebraically (and not exponentially).

We tested the effectiveness of X-REGO numerically with three global and one local solver on the set of functions with low effective dimensionality. We proposed different variants of X-REGO each corresponding to a specific rule for choosing $\boldsymbol{p}$'s and contrasted them against each other and against the 'no-embedding' framework in which the solvers were applied to (P) directly with no use of subspace embeddings. The results of the experiments showed that the difference in performance between X-REGO and 'no-embedding' became more prominent for larger D in favour of X-REGO. The results further suggest that the effectiveness of X-REGO (just like of REGO) is solver-dependent. In our experiments, the best results were achieved by the local solver.

In the future, we may consider further exploring the numerical and theoretical potential of the X-REGO algorithmic framework. In particular, we may apply X-REGO for functions with low effective dimensionality, setting the embedding dimension d to be less than the effective subspace dimension d_e (for instance, we may progressively increase d starting from 1 until we observe no improvement in the solutions of subproblems). We anticipate that the results of Chapter 5 may be useful in securing the convergence of X-REGO for functions with low effective dimensionality in the case when d_e is unknown. Finally, we plan to investigate the performance of X-REGO when applied to general objectives and compare it popular global optimization solvers.

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover, New York, ninth dover printing, tenth gpo printing edition, 1964.

[2] C.S. Adjiman, I.P. Androulakis, C.D. Maranas, and C.A. Floudas. A global optimization method, αbb, for process design. *Computers & Chemical Engineering*, 20:S419 – S424, 1996. European Symposium on Computer Aided Process Engineering-6.

[3] T. Amdeberhan and V. H. Moll, editors. *Tapas in Experimental Mathematics*, Contemporary Mathematics 457, 2008. American Mathematical Society.

[4] D. Amelunxen. *Geometric analysis of the condition of the convex feasibility problem.* PhD thesis, University of Paderborn, 2011.

[5] D. Amelunxen and M. Lotz. Intrinsic volumes of polyhedral cones: A combinatorial perspective. *Discrete & Computational Geometry*, 58(2):371–409, 2017.

[6] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

[7] I. P. Androulakis, C. D. Maranas, and C. A. Floudas. αbb: A global optimization method for general constrained nonconvex problems. *Journal of Global Optimization*, 7(4):337–363, 1995.

[8] C. Audet and D. Orban. Finding optimal algorithmic parameters using derivative-free optimization. *SIAM Journal on Optimization*, 17(3):642–664, 2006.

[9] F. Bagattini, F. Schoen, and L. Tigli. Clustering methods for the optimization of atomic cluster structure. *The Journal of Chemical Physics*, 148(14):144102, 2018.

[10] F. Bagattini, F. Schoen, and L. Tigli. Clustering methods for large scale geometrical global optimization. *Optimization Methods and Software*, 34(5):1099–1122, 2019.

[11] E. Balas, F. Glover, and S. Zionts. An additive algorithm for solving linear programs with zero-one variables. *Operations Research*, 13(4):517–549, 1965.

[12] A. S. Bandeira, K. Scheinberg, and L. N. Vicente. Convergence of trust-region methods based on probabilistic models. *SIAM Journal on Optimization*, 24(3):1238–1264, 2014.

[13] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(4):2037–2060, 2013.

[14] R.W. Becker and G.V. Lago. A global optimization algorithm. In *Prceedings of the 8th Allerton Conference on Cicuits and Systems Theory*, pages 3–12, Monticello, Illinois, 1970.

[15] M. Ben Salem, F. Bachoc, O. Roustant, F. Gamboa, and L. Tomaso. Sequential dimension reduction for learning features of expensive black-box functions. hal-01688329v2, 2019.

[16] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization.* Society for Industrial and Applied Mathematics, 2001.

[17] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(1):281–305, 2012.

[18] J. M. Bernardo and A. F. M. Smith. *Bayesian theory.* Wiley, 2000.

[19] M. Binois, D. Ginsbourger, and O. Roustant. A warped kernel improving robustness in bayesian optimization via random embeddings. In *Learning and Intelligent Optimization*, pages 281–286, Cham, 2015. Springer International Publishing.

[20] M. Binois, D. Ginsbourger, and O. Roustant. On the choice of the low-dimensional domain for global optimization via random embeddings. *Journal of Global Optimization*, 76(1):69–90, 2020.

[21] A. Booker, P. Frank, Jr. J. Dennis, D. Moore, and D. Serafini. *Managing surrogate objectives to optimize a helicopter rotor design - Further experiments.* 1998.

[22] A. J. Booker, J. E. Dennis, P. D. Frank, D. B. Serafini, and V. Torczon. *Optimization Using Surrogate Objectives on a Helicopter Test Example*, pages 49–58. Birkhäuser Boston, Boston, MA, 1998.

[23] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, New York, NY, USA, 2004.

[24] E. Brochu, V. M. Cora, and N. de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning, 2009. Technical Report TR-2009-23, University of British Columbia, Department of Computer Science.

[25] E. Brochu, T. Brochu, and N. de Freitas. A bayesian interactive optimization approach to procedural animation design. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '10, pages 103–112, Goslar, DEU, 2010. Eurographics Association.

[26] S. H. Brooks. A discussion of random methods for seeking maxima. *Operations Research*, 6(2):244–251, 1958.

[27] R. L. Burden and J. D. Faires. *Numerical Analysis.* The Prindle, Weber and Schmidt Series in Mathematics. PWS-Kent Publishing Company, Boston, fourth edition, 1989.

[28] R. H. Byrd, J. Nocedal, and R. A. Waltz. *Knitro: An Integrated Package for Nonlinear Optimization*, pages 35–59. Springer US, Boston, MA, 2006.

[29] C. Cartis and A. Otemissov. A dimensionality reduction technique for unconstrained global optimization of functions with low effective dimensionality. *arXiv e-prints*, page arXiv:2003.09673, 2020. To appear in *Information and Inference: a journal of the IMA.*

[30] C. Cartis and K. Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Math. Program.*, 169(2):337–375, 2018.

[31] C. Cartis, T. Ferguson, and L. Roberts. Scalable Derivative-Free Optimization for Nonlinear Least-Squares Problems. *arXiv e-prints*, art. arXiv:2007.13243, 2020.

[32] C. Cartis, E. Massart, and A. Otemissov. Constrained global optimization of functions with low effective dimensionality using multiple random embeddings. *arXiv e-prints*, art. arXiv:2009.10446, 2020.

[33] B. Chen, A. Krause, and R. M. Castro. Joint optimization and variable selection of high-dimensional gaussian processes. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1423–1430, 2012.

[34] J. Chen, G. Zhu, R. Gu, C. Yuan, and Y. Huang. Semi-supervised embedding learning for high-dimensional bayesian optimization. *arXiv e-prints*, page arXiv:2005.14601, 2020.

[35] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust Region Methods.* Society for Industrial and Applied Mathematics, 2000.

[36] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to Derivative-Free Optimization.* Society for Industrial and Applied Mathematics, 2009.

[37] P. Constantine. *Active Subspaces.* SIAM, Philadelphia, PA, 2015.

[38] G. Dantzig. On the significance of solving linear programming problems with some integer variables. *Econometrica*, 28(1):30–44, 1960.

[39] G. Dantzig, S. Johnson, and W. White. A linear programming approach to the chemical equilibrium problem. *Management Science*, 5(1):38–43, 1958.

[40] K. R. Davidson and S. Szarek. Local operator theory, random matrices and banach spaces. *Handbook on the Geometry of Banach spaces, Vol. 1*, pages 317–366, 2001.

[41] N. Demo, M. Tezzele, and G. Rozza. A supervised learning approach involving active subspaces for an efficient genetic algorithm in high-dimensional optimization problems. *arXiv e-prints*, page arXiv:2006.07282, 2020.

[42] L. C. W. Dixon and G. P. Szegö. *Towards Global Optimization.* Elseiver, New York, 1975.

[43] J. Djolonga, A. Krause, and V. Cevher. High-dimensional gaussian process bandits. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 1025–1033, 2013.

[44] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, 2002.

[45] N. J. Driebeek. An algorithm for the solution of mixed integer programming problems. *Management Science*, 12(7):576–587, 1966.

[46] R. Durrett. *Probability: Theory and Examples.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition, 2019.

[47] A. Edelman. Eigenvalues and condition numbers of random matrices. *SIAM Journal on Matrix Analysis and Applications*, 9(4):543–560, 1988.

[48] D. Eriksson, K. Dong, E. H. Lee, D. Bindel, and A. G. Wilson. Scaling gaussian process regression with derivatives. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 6868–6878, 2018.

[49] P. A. Ernesto and U. P. Diliman. MVF—multivariate test functions library in C for unconstrained global optimization, 2005.

[50] J. E. Falk and R. M. Soland. An algorithm for separable nonconvex programming problems. *Management Science*, 15(9):550–569, 1969.

[51] K. Fang, S. Kotz, and K. W. Ng. *Symmetric multivariate and related distributions.* London: Chapman and Hall, 1990.

[52] D. E. Finkel. *Direct optimization algorithm user guide.* North Carolina State University, 2003. Available at http://www2.peq.coppe.ufrj.br/Pessoal/Professores/Arge/COQ897/Naturais/DirectUserGuide.pdf.

[53] R. Flamary, A. Rakotomamonjy, and G. Gasso. Importance sampling strategy for non-convex randomized block-coordinate descent. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 301–304, 2015.

[54] C. A. Floudas. *Deterministic Global Optimization: Theory, Methods and Applications.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[55] M. Fornasier, K. Schnass, and J. Vybiral. Learning functions of few arbitrary linear parameters in high dimensions. *Foundations of Computational Mathematics*, 12(2): 229–262, 2012.

[56] P. I. Frazier. A Tutorial on Bayesian Optimization. *arXiv e-prints*, art. arXiv:1807.02811, July 2018.

[57] L. P. Fröhlich, E. D. Klenske, C. G. Daniel, and M. N. Zeilinger. Bayesian optimization for policy search in high-dimensional systems via automatic domain selection. *arXiv e-prints*, 2020.

[58] J. M. Gablonsky and C. T. Kelley. A locally-biased form of the direct algorithm. *Journal of Global Optimization*, 21(1):27–37, 2001.

[59] E. A. Galperin. The cubic algorithm. *Journal of Mathematical Analysis and Applications*, 112(2):635–640, 1985.

[60] R. Garnett, M. A. Osborne, and P. Hennig. Active learning of linear embeddings for gaussian processes. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI'14, pages 230–239, 2014.

[61] C.-Y. Gau and L. E. Schrage. Implementation and testing of a branch-and-bound based method for deterministic global optimization: Operations research applications. In C. A. Floudas and Panos Pardalos, editors, *Frontiers in Global Optimization*, pages 145–164. Springer US, Boston, MA, 2004.

[62] A. Gavana. Global optimization benchmarks and AMPGO. Available at http://infinity77.net/global_optimization/.

[63] P. E. Gill, W. Murray, and M. A. Saunders. Snopt: An sqp algorithm for large-scale constrained optimization. *SIAM Review*, 47(1):99–131, 2005.

[64] P. E. Gill, W. Murray, and Saunders M. A. *User's Guide for SNOPT Version 7: Software for Large-Scale Nonlinear Programming.* Stanford University and University of California, San Diego, 2008. Available at https://web.stanford.edu/group/SOL/guides/sndoc7.pdf.

[65] A. Giovanoglou, A. Galindo, G. Jackson, and C.S. Adjiman. Fluid phase stability and equilibrium calculations in binary mixtures: Part i: Theoretical development for non-azeotropic mixtures. *Fluid Phase Equilibria*, 275(2):79–94, 2009.

[66] L. Goldstein, I. Nourdin, and G. Peccati. Gaussian phase transitions and conic intrinsic volumes: Steining the steiner formula. *The Annals of Applied Probability*, 27 (1):1–47, 2017.

[67] R. E. Gomory. Outline of an algorithm for integer solutions to linear programs. *Bulletin of the American Mathematical Society*, 64(5):275–278, 1958.

[68] E. Gorbunov, F. Hanzely, and P. Richtárik. A Unified Theory of SGD: Variance Reduction, Sampling, Quantization and Coordinate Descent. *arXiv e-prints*, art. arXiv:1905.11261, 2019.

[69] N. I. M. Gould, D. Orban, and P. L. Toint. Galahad, a library of thread-safe fortran 90 packages for large-scale nonlinear optimization. *ACM Trans. Math. Softw.*, 29(4): 353–372, 2003.

[70] A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions.* New York: Chapman and Hall/CRC, 2000.

[71] A. K. Gupta and D. Song. Lp-norm spherical distribution. *Journal of Statistical Planning and Inference*, 60(2):241–260, 1997.

[72] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[73] J. H. Holland. Genetic algorithms and the optimal allocation of trials. *SIAM Journal on Computing*, 2(2):88–105, 1973.

[74] R. Horst and P. M. Pardalos, editors. *Handbook of Global Optimization*, volume 1. Springer-Verlag New York, Inc., 1995.

[75] R. Horst and N. V. Thoai. DC programming: Overview. *Journal of Optimization Theory and Applications*, 103(1):1–43, 1999.

[76] F. Hutter, H. Hoos, and K. Leyton-Brown. An efficient approach for assessing hyperparameter importance. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages I–754–I–762, 2014.

[77] D. R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21(4):345–383, Dec 2001.

[78] D. R. Jones, C. D. Perttunen, and B. E. Stuckman. Lipschitzian optimization without the Lipschitz constant. *Journal of Optimization Theory and Applications*, 79(1):157–181, 1993.

[79] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[80] A. H. G. Rinnooy Kan and G. T. Timmer. Chapter IX Global optimization. In *Optimization*, volume 1 of *Handbooks in Operations Research and Management Science*, pages 631 – 662. Elsevier, 1989.

[81] K. Kandasamy, J. Schneider, and B. Póczos. High dimensional bayesian optimisation and bandits via additive models. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 295–304, 2015.

[82] R. B. Kearfott. *GlobSol: History, Composition, and Advice on Use*, pages 17–31. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[83] C. T. Kelley. *Iterative Methods for Optimization.* Society for Industrial and Applied Mathematics, 1999.

[84] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671, 1983.

[85] J. Kirschner, M. Mutny, N. Hiller, R. Ischebeck, and A. Krause. Adaptive and safe Bayesian optimization in high dimensions via one-dimensional subspaces. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3429–3438, 2019.

[86] C. G. Knight, S. H. E. Knight, N. Massey, T. Aina, C. Christensen, D. J. Frame, J. A. Kettleborough, A. Martin, S. Pascoe, B. Sanderson, D. A. Stainforth, and M. R. Allen. Association of parameter, software, and hardware variation with large-scale behavior across 57,000 climate models. *Proceedings of the National Academy of Sciences*, 104 (30):12259–12264, 2007.

[87] T. G. Kolda, R. M. Lewis, and V. Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review*, 45(3):385–482, 2003.

[88] P. Kolesar. A branch and bound algorithm for the knapsack problem. *Management Science*, 13:723–735, 1967.

[89] H. J. Kushner. A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise. *Journal of Basic Engineering*, 86(1):97–106, 1964.

[90] A. H. Land and A. G. Doig. An automatic method of solving discrete programming problems. *Econometrica*, 28(3):497–520, 1960.

[91] J. Larson, M. Menickelly, and S. M. Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019.

[92] P. M. Lee. *Bayesian Statistics: An Introduction.* John Wiley & Sons, 4th edition, 2012.

[93] B. Letham, R. Calandra, A. Rai, and E. Bakshy. Re-Examining Linear Embeddings for High-Dimensional Bayesian Optimization. *arXiv e-prints*, art. arXiv:2001.11659, 2020.

[94] T. Levina, Y. Levin, J. McGill, and M. Nediak. Dynamic pricing with online learning and strategic consumers: An application of the aggregating algorithm. *Operations Research*, 57(2):327–341, 2009.

[95] A. V. Levy and S. Gomez. The tunneling algorithm for the global optimization problem of constrained functions. Technical report, Universidad Nacional Autonoma de Mexico, 1980.

[96] A. V. Levy and A. Montalvo. The tunneling algorithm for the global minimization of functions. *SIAM Journal on Scientific and Statistical Computing*, 6(1):15–29, 1985.

[97] C.-L. Li, K. Kandasamy, B. Poczos, and J. Schneider. High dimensional bayesian optimization via restricted projection pursuit models. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 884–892, 2016.

[98] L. Liberti and S. Kucherenko. Comparison of deterministic and stochastic approaches to global optimization. *International Transactions in Operational Research*, 12(3): 263–285, 2005.

[99] J. D. C. Little, K. G. Murty, D. W. Sweeney, and C. Karel. An algorithm for the traveling salesman problem. *Operations Research*, 11(6):972–989, 1963.

[100] D. Lizotte, T. Wang, M. Bowling, and D. Schuurmans. Automatic gait optimization with gaussian process regression. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 944–949, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[101] M. Locatelli and F. Schoen. *Global Optimization: Theory, Algorithms and Applications.* Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013.

[102] R. Mansel Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtarik. SGD: General Analysis and Improved Rates. *arXiv e-prints*, art. arXiv:1901.09401, 2019.

[103] H. M. Markowitz and A. S. Manne. On the solution of discrete programming problems. *Econometrica*, 25(1):84–110, 1957.

[104] M. B. McCoy and J. A. Tropp. From steiner formulas for cones to concentration of intrinsic volumes. *Discrete & Computational Geometry*, 51(4):926–963, 2014.

[105] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[106] R. Misener and C. A. Floudas. ANTIGONE: Algorithms for coNTinuous / Integer Global Optimization of Nonlinear Equations. *Journal of Global Optimization*, 59(2): 503–526, 2014.

[107] L. G. Mitten. Branch-and-bound methods: General formulation and properties. *Operations Research*, 18(1):24–34, 1970.

[108] J. Močkus. On bayesian methods for seeking the extremum. In G. I. Marchuk, editor, *Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974*, pages 400–404, Berlin, Heidelberg, 1975. Springer Berlin Heidelberg.

[109] J. Močkus. The bayesian approach to global optimization. In R. F. Drenick and F. Kozin, editors, *System Modeling and Optimization*, pages 473–481. Springer, Berlin, Heidelberg, 1982.

[110] J. Močkus. *Bayesian approach to global optimization: theory and applications.* Dordrecht : Kluwer Academic, Boston, 1989.

[111] J. Močkus, V. Tiesis, and A. Zhilinskas. The application of bayesian methods for seeking the extremum. volume 2, pages 117–129. 09 1978.

[112] R. E. Moore. *Interval Arithmetic and Automatic Error Analysis in Digital Computing.* PhD thesis, Stanford, CA, USA, 1963. AAI6304614.

[113] R. E. Moore and C. T. Yang. Interval analysis. Technical report, Lockheed Missiles and Space Co., 1959. I. Space Division Report LMSD285875.

[114] A. Nayebi, A. Munteanu, and M. Poloczek. A framework for Bayesian optimization in embedded subspaces. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4752–4761, 2019.

[115] D. M. Negoescu, P. I. Frazier, and W. B. Powell. The knowledge-gradient algorithm for sequencing experiments in drug discovery. *INFORMS J. on Computing*, 23(3): 346–363, 2011.

[116] J. A. Nelder and R. Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 1965.

[117] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming.* Society for Industrial and Applied Mathematics, 1994.

[118] A. Neumaier. Complete search in continuous global optimization and constraint satisfaction. *Acta Numerica*, 13:271–369, 2004.

[119] A. Neumaier, O. Shcherbina, W. Huyer, and T. Vinkó. A comparison of complete global optimization solvers. *Mathematical Programming*, 103(2):335–356, 2005.

[120] E. Neuman. Inequalities and bounds for the incomplete gamma function. *Results in Mathematics*, 63(3):1209–1214, 2013.

[121] NIST. NIST Digital Library of Mathematical Functions, 2020. Available at https://dlmf.nist.gov.

[122] NIST/SEMATECH. e-handbook of statistical methods, 2018. Available at https://www.itl.nist.gov/div898/handbook/eda/section3/eda3666.htm.

[123] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer, New York, NY, USA, second edition, 2006.

[124] S. Oymak and J. A. Tropp. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 7(3):337–446, 2017.

[125] N. R. Patel, R. L. Smith, and Z. B. Zabinsky. Pure adaptive search in monte carlo optimization. *Mathematical Programming*, 43(1):317–328, 1989.

[126] F. E. Pereira, G. Jackson, A. Galindo, and C. S. Adjiman. A duality-based optimisation approach for the reliable solution of (p, t) phase equilibrium in volume-composition space. *Fluid Phase Equilibria*, 299(1):1–23, 2010.

[127] J. D. Pintér. Globally convergent methods for n-dimensional multiextremal optimization. *Optimization*, 17(2):187–202, 1986.

[128] J. D. Pintér. *Global Optimization in Action: Continuous and Lipschitz Optimization: Algorithms, Implementations and Applications.* Kluwer Academic Publishers, 1996.

[129] S. A. Piyavskii. An algorithm for finding the absolute extremum of a function. *USSR Computational Mathematics and Mathematical Physics*, 12(4):57 – 67, 1972.

[130] M. H. Protter and C. B. Jr. Morrey. *A First Course in Real Analysis.* Springer, New York, 1977.

[131] H. Qian and Y. Yu. Solving high-dimensional multi-objective optimization problems with low effective dimensions. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pages 875–881, 2017.

[132] H. Qian, Y.-Q. Hu, and Y. Yu. Derivative-free optimization of high-dimensional non-convex functions by sequential random embeddings. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 1946–1952, 2016.

[133] G. Raskutti and M. Mahoney. A Statistical Perspective on Randomized Sketching for Ordinary Least-Squares. *arXiv e-prints*, art. arXiv:1406.5986, 2014.

[134] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning).* The MIT Press, 2005.

[135] L. Roberts. *Derivative-free algorithms for nonlinear optimisation problems.* PhD thesis, University of Oxford, 2019.

[136] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

[137] P. Rolland, J. Scarlett, I. Bogunovic, and V. Cevher. High-Dimensional Bayesian Optimization via Additive Models with Overlapping Groups. *arXiv e-prints*, page arXiv:1802.07028, 2018.

[138] M. Rudelson and R. Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009.

[139] N. V. Sahinidis. *BARON 14.3.1: Global Optimization of Mixed-Integer Nonlinear Programs,* User's Manual, 2014.

[140] A. Sankar, D. A. Spielman, and S.-H. Teng. Smoothed analysis of the condition numbers and growth factors of matrices. *SIAM Journal on Matrix Analysis and Applications*, 28(2):446–476, 2006.

[141] M. L. Sanyang and A. Kabán. Remeda: Random embedding eda for optimising functions with intrinsic dimension. In *Parallel Problem Solving from Nature – PPSN XIV*, pages 859–868, 2016.

[142] R. Schneider and W. Weil. *Stochastic and Integral Geometry.* Springer series in statistics: Probability and its applications. Springer, 2008.

[143] B. O. Shubert. A sequential method seeking the global maximum of a function. *SIAM Journal on Numerical Analysis*, 9(3):379–10, 09 1972.

[144] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'12, page 2951–2959, Red Hook, NY, USA, 2012. Curran Associates Inc.

[145] F. J. Solis and R. J.-B. Wets. Minimization by random search techniques. *Mathematics of Operations Research*, 6(1):19–30, 1981.

[146] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 1015–1022, Madison, WI, USA, 2010.

[147] S. Surjanovic and D. Bingham. Virtual library of simulation experiments: Test functions and datasets, 2013. Available at https://www.sfu.ca/~ssurjano/.

[148] M. Tawarmalani and N. V. Sahinidis. A polyhedral branch-and-cut approach to global optimization. *Mathematical Programming*, 103:225–249, 2005.

[149] N. Temme. *Asymptotic Methods for Integrals.* World Scientific, Singapore, 2014.

[150] S. F. B. Tett, M. J. Mineter, C. Cartis, D. J. Rowlands, and P. Liu. Can Top-of-Atmosphere Radiation Measurements Constrain Climate Predictions? Part I: Tuning. *Journal of Climate*, 26(23):9348–9366, 2013.

[151] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[152] A. A. Törn. Global optimization as a combination of local and global search. In W. Goldberg, editor, *Computer Simulation Versus Analytical Solutions for Business and Economical Models*, Gothenburg, 1972".

[153] A. A. Törn. *Global optimization as a combination of global and local search.* PhD thesis, Abo Akademi University, 1974. HHAAA 13.

[154] A. A. Törn. A search clustering approach to global optimization. In *Towards Global Optimization*, volume 2. 1978.

[155] F. G. Tricomi and A. Erdélyi. The asymptotic expansion of a ratio of gamma functions. *Pacific J. Math.*, 1:133–142, 1951.

[156] H. Tyagi and V. Cevher. Learning non-parametric basis independent models from point queries via low-rank methods. *Applied and Computational Harmonic Analysis*, 37(3):389–412, 2014.

[157] P. J. M. van Laarhoven and E. H. L. Aarts. *Simulated annealing : theory and applications.* Mathematics and its applications. Dordrecht, 1987.

[158] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, 2012.

[159] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge Series in Statistical and Probabilistic Mathematics. 2018.

[160] A.V. Vilkov, N.P. Zhidkov, and B.M. Shchedrin. A method of finding the global minimum of a function of one variable. *USSR Computational Mathematics and Mathematical Physics*, 15(4):221–224, 1975.

[161] A. G. Žilinskas. Single-step bayesian search method for an extremum of functions of a single variable. *Cybernetics*, 11(1):160–166, 1975.

[162] A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.

[163] Z. Wang, F. Hutter, M. Zoghi, D. Matheson, and N. De Freitas. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55(1):361–387, 2016.

[164] Z. Wang, C. Gehring, P. Kohli, and S. Jegelka. Batched large-scale bayesian optimization in high-dimensional spaces. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

[165] R. L. Wheeden. *Measure and integral : an introduction to real analysis.* Boca Raton: Chapman and Hall/CRC, 2nd edition, 2015.

[166] R. Wong. *Asymptotic Approximations of Integrals.* Society for Industrial and Applied Mathematics, 2001.

[167] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[168] S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1): 3–34, 2015.

[169] Y. Xu and W. Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013.

[170] Y.-T. Xu, Y. Zhang, and S.-G. Wang. A modified tunneling function method for non-smooth global optimization and its application in artificial neural network. *Applied Mathematical Modelling*, 39(21):6438–6450, 2015.

[171] R. Yang. Convergence of the simulated annealing algorithm for continuous global optimization. *Journal of Optimization Theory and Applications*, 104(3):691–716, 2000.

[172] Z. B. Zabinsky. *Stochastic Adaptive Search for Global Optimization.* Springer Science & Business Media, 2003.

[173] Z.B. Zabinsky and R.l. Smith. Pure adaptive search in global optimization. *Mathematical Programming*, 53(3):323–338, 1992.

[174] M. Zhang, H. Li, and S. Su. High dimensional bayesian optimization via supervised dimension reduction. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, IJCAI'19, pages 4292–4298, 2019.

Appendices

Appendix A

Technical results and Proofs

A.1 Extreme singular values of Gaussian matrices

Gaussian matrices are random and so are their singular values. We write $s_{\min}(\boldsymbol{M})$ and $s_{\max}(\boldsymbol{M})$ to refer to the smallest and the largest singular values of $\boldsymbol{M}$. Many probabilistic bounds have been derived for the tails of the extreme singular values of Gaussian matrices. Below we state some well-known non-asymptotic bounds.

Theorem A.1. *(see [47], [140, Theorem 3.4]) Let* $\boldsymbol{M}$ *be an* $M_1 \times M_1$ *Gaussian random matrix. Then, we have*

$$\mathbb{P}[s_{\min}(\boldsymbol{M}) \leq \epsilon] \leq \sqrt{M_1}\epsilon.$$

Theorem A.2. *(see [40], [158, Corollary 5.35]) Let* $\boldsymbol{M}$ *be an* $M_1 \times M_2$ *Gaussian random matrix* $M_1 \geq M_2$. *Then, we have*

$$\mathbb{P}[\sqrt{M_1} - \sqrt{M_2} - t \leq s_{\min}(\boldsymbol{M}) \leq s_{\max}(\boldsymbol{M}) \leq \sqrt{M_1} + \sqrt{M_2} + t] \geq 1 - 2e^{-t^2/2}$$

for any $t \geq 0$.

Theorem A.3. *(see [138, Theorem 1.1]) Let* $\boldsymbol{M}$ *be an* $M_1 \times M_2$ *Gaussian random matrix,* $M_1 \geq M_2$. *Then, for every* $\epsilon > 0$, *we have*

$$\mathbb{P}[s_{\min}(\boldsymbol{M}) \leq \epsilon(\sqrt{M_1} - \sqrt{M_2 - 1})] \leq (C\epsilon)^{M_1 - M_2 + 1} + e^{-cM_1},$$

where $C, c > 0$ *are absolute constants.*

Theorem A.3 is a corollary of a more general result in [138], which is applicable for matrices with sub-Gaussian entries.

A.2 Miscellaneous probability results

In this section, we provide three probability results that we use in the proofs of theorems in Chapters 2 and 4.

First, we prove two short lemmas that we use in Chapter 2. The first lemma derives an upper bound for the cumulative density function of a chi-squared random variable. To prove the first lemma, we need the following upper bound on the lower incomplete gamma function $\gamma(a,t)$ defined in (2.3).

Theorem A.4. *(see [120, Theorem 4.1]) Let $\gamma(a,t)$ be the lower incomplete gamma function defined in* (2.3). *Then,*

$$\gamma(a,t) \leq \frac{1}{a(a+1)}(1+ae^{-t})t^a. \tag{A.1}$$

We state our first lemma.

Lemma A.5. *Let $X \sim \chi_N^2$. Then,*

$$\mathbb{P}[X \leq x] \leq \frac{4}{N(N+2)\Gamma(N/2)}\left(1+\frac{N}{2}e^{-x/2}\right)(x/2)^{N/2}. \tag{A.2}$$

Proof. Recall the c.d.f. of the chi-square random variable in (2.1):

$$\mathbb{P}[X \leq x] = \frac{\gamma(N/2, x/2)}{\Gamma(N/2)} \tag{A.3}$$

for $x > 0$. We obtain the desired result by applying (A.1) to upper bound $\gamma(N/2, x/2)$ in (A.3). ☐

In the second lemma, we derive the p.d.f. of a random variable that is a square root of an inverse chi-squared random variable.

Lemma A.6. *Let Y and R be two random variables such that $Y \sim 1/\chi_N^2$ and $R = \sqrt{Y}$. Then, the probability density function (p.d.f.) $g(\hat{r})$ of R is given by*

$$g(\hat{r}) = \frac{2^{-N/2+1}}{\Gamma(N/2)}\hat{r}^{-N-1}e^{-1/(2\hat{r}^2)} \text{ for } \hat{r} > 0. \tag{A.4}$$

Proof. The p.d.f. $g(\hat{r})$ of R satisfies

$$g(\hat{r}) = \frac{d}{d\hat{r}}\,\mathbb{P}[\sqrt{Y} < \hat{r}] = \frac{d}{d\hat{r}}\,\mathbb{P}[Y < \hat{r}^2] = 2\hat{r}h(N, \hat{r}^2), \tag{A.5}$$

where $h(N,\cdot)$ is the p.d.f. of Y. By substituting the formula for $h(N,\cdot)$ provided in (2.4) in the above, we obtain the desired result. ☐

Lastly, we provide a useful known result in probability theory, which says that if two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ follow the same distribution then $f(\boldsymbol{x})$ and $f(\boldsymbol{y})$ follow the same distribution if the function f is measurable.

Lemma A.7. *[51, p. 13] Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be random vectors such that $\boldsymbol{x} \overset{law}{=} \boldsymbol{y}$ and let $f_i(\cdot)$, $i = 1, 2, \ldots, m$, be measurable functions. Then,*

$$\begin{pmatrix} f_1(\boldsymbol{x}) & f_2(\boldsymbol{x}) & \ldots & f_m(\boldsymbol{x}) \end{pmatrix}^T \overset{law}{=} \begin{pmatrix} f_1(\boldsymbol{y}) & f_2(\boldsymbol{y}) & \ldots & f_m(\boldsymbol{y}) \end{pmatrix}^T.$$

A.3 Proofs of Section 4.2.3

This section derives a few results that are used in Section 4.2.3.

A.3.1 Properties of $J_{m,n}(\Delta)$

Here, we establish two properties of the integral $J_{m,n}(\Delta)$ defined in (4.36). First, we show that $J_{m,n}(\Delta)$ is a monotonically decreasing function of Δ.

Lemma A.8. *The integral $J_{m,n}(\Delta)$ in* (4.36) *is a monotonically decreasing function of Δ.*

Proof. Let Δ_1, Δ_2 be any positive reals that satisfy $\Delta_1 \leq \Delta_2$. We need to show that $J_{m,n}(\Delta_1) \geq J_{m,n}(\Delta_2)$. This relation follows immediately from the observation that, for any $s \geq 0$,

$$\sqrt{\frac{2}{\pi}} \int_0^{s/\Delta_1} e^{-x^2/2} dx \geq \sqrt{\frac{2}{\pi}} \int_0^{s/\Delta_2} e^{-x^2/2} dx$$

since the integrand is positive. □

Next, we show that $J_{m,n}(\Delta)$ cannot be larger than 1 if $\Delta > 0$.

Lemma A.9. *The integral $J_{m,n}(\Delta)$ defined in* (4.36) *satisfies $J_{m,n}(\Delta) \leq 1$ for all $\Delta > 0$.*

Proof. Note that, for any $s \geq 0$, we have

$$\sqrt{\frac{2}{\pi}} \int_0^{s/\Delta} e^{-x^2/2} dx \leq \sqrt{\frac{2}{\pi}} \int_0^{\infty} e^{-x^2/2} dx = 1.$$

Hence,

$$J_{m,n}(\Delta) \leq \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^{\infty} s^{n-1} e^{-s^2/2} ds = 1.$$ □

A.3.2 Asymptotic expansions of two integrals

Let $L(\lambda, \mu, z)$ and $G(\lambda, \mu, z)$ be defined in (4.69) and (4.70), respectively. In the following two theorems we derive the asymptotic expansions of $L(\lambda, \mu, z)$ and $G(\lambda, \mu, z)$ for large z. These results are needed to assist the proof of Theorem 4.20.

First, we derive the asymptotic expansion for $G(\lambda, \mu, z)$ closely following the proof in [166, Theorem 2, p. 70].

Theorem A.10. *Let* $0 < c < 1$ *and let* λ *and* μ *be any real numbers with* $\lambda > 0$. *We have*

$$\begin{aligned} G(\lambda, \mu, z) \sim & z^{-\lambda}(\log(z))^{\mu} \log(\log(z)) \sum_{r=0}^{\infty} (-1)^r \binom{\mu}{r} \Gamma^{(r)}(\lambda)(\log(z))^{-r} + \\ & + z^{-\lambda}(\log(z))^{\mu} \sum_{r=1}^{\infty} a_r \Gamma^{(r)}(\lambda)(\log(z))^{-r}, \end{aligned}$$

as $z \to \infty$, *where* $\binom{\mu}{i}$ *is defined in* (5.14), $\Gamma^{(r)}$ *denotes the* r*th derivative of the gamma function, and where*

$$a_r = -\sum_{i=0}^{r-1} \binom{\mu}{i} \frac{(-1)^i}{r-i} \text{ for } r = 1, 2, \ldots. \tag{A.6}$$

Proof. With the substitution $u = zt$, we obtain

$$\begin{aligned} G(\lambda, \mu, z) &= z^{-\lambda} \int_0^{cz} u^{\lambda-1}(\log(z) - \log(u))^{\mu} \log(\log(z) - \log(u)) e^{-u} du \\ &= z^{-\lambda}(\log(z))^{\mu} \int_0^{cz} u^{\lambda-1} \left(1 - \frac{\log(u)}{\log(z)}\right)^{\mu} \left(\log(\log(z)) + \log\left(1 - \frac{\log(u)}{\log(z)}\right)\right) e^{-u} du \\ &= z^{-\lambda}(\log(z))^{\mu}(\log(\log(z))G_1 + G_2), \end{aligned} \tag{A.7}$$

where

$$G_1 = \int_0^{cz} u^{\lambda-1} \left(1 - \frac{\log(u)}{\log(z)}\right)^{\mu} e^{-u} du$$

and

$$G_2 = \int_0^{cz} u^{\lambda-1} \left(1 - \frac{\log(u)}{\log(z)}\right)^{\mu} \log\left(1 - \frac{\log(u)}{\log(z)}\right) e^{-u} du. \tag{A.8}$$

We first derive the asymptotic expansion for G_2, the asymptotic expansion for G_1 can then be derived in a similar manner.

Let N be an arbitrary positive integer such that $N + 1 \geq \mu$. By Taylor's expansion,

$$\begin{aligned} \left(1 - \frac{\log(u)}{\log(z)}\right)^{\mu} &= \sum_{r=0}^{N} (-1)^r \binom{\mu}{r} \left(\frac{\log(u)}{\log(z)}\right)^r + R_{1,N} \\ \log\left(1 - \frac{\log(u)}{\log(z)}\right) &= -\sum_{r=1}^{N} \frac{1}{r} \left(\frac{\log(u)}{\log(z)}\right)^r + R_{2,N}, \end{aligned}$$

for all $0 < u < cz$, where

$$|R_{i,N}| \leq C_{i,N} \frac{|\log(u)|^{N+1}}{|\log(z)|^{N+1}} \ (i = 1, 2)$$

for some fixed constants $C_{1,N}, C_{2,N} > 0$. Hence,

$$\left(1 - \frac{\log(u)}{\log(z)}\right)^{\mu} \log\left(1 - \frac{\log(u)}{\log(z)}\right) = \sum_{r=1}^{2N} a_r \left(\frac{\log(u)}{\log(z)}\right)^{r} + R_{2N}, \tag{A.9}$$

for all $0 < u < cz$, where a_r's are defined as in (A.6) and

$$|R_{2N}| \leq C_{2N} \frac{|\log(u)|^{2N+1}}{|\log(z)|^{2N+1}}$$

for some fixed $C_{2N} > 0$. By substituting (A.9) in (A.8), we obtain

$$G_2 = \sum_{r=1}^{2N} a_r (\log(z))^{-r} \int_0^{cz} u^{\lambda-1} (\log(u))^r e^{-u} du + r_{2N},$$

where

$$r_{2N} = \int_0^{cz} u^{\lambda-1} e^{-u} R_{2N} du.$$

Wong showed in [166, p. 71] that, as $z \to \infty$,

$$\int_0^{cz} u^{\lambda-1} (\log(u))^r e^{-u} du = \Gamma^{(r)}(\lambda) + O(e^{-\epsilon cz}),$$

where $\epsilon \in (0, 1/2)$. Furthermore,

$$\begin{aligned} |r_{2N}| &\leq C_{2N} |\log(z)|^{-2N-1} \int_0^{cz} |u^{\lambda-1} \log(u)^{2N+1} e^{-u}| du \\ &\leq C_{2N} |\log(z)|^{-2N-1} \int_0^{\infty} |u^{\lambda-1} \log(u)^{2N+1} e^{-u}| du \end{aligned}$$

It can be shown that the latter integral is bounded (see [166, eq. (2.27), p. 71]; thus, $r_{2N} = O(\log(z)^{-2N-1})$. Hence,

$$G_2 = \sum_{r=1}^{2N} a_r \Gamma^{(r)}(\lambda) (\log(z))^{-r} + O(\log(z)^{-2N-1}). \tag{A.10}$$

In a similar manner, one can show that

$$G_1 = \sum_{r=0}^{N} (-1)^r \binom{\mu}{r} \Gamma^{(r)}(\lambda) (\log(z))^{-r} + O(\log(z)^{-N-1}). \tag{A.11}$$

Combining (A.7), (A.10) and (A.11), we obtain the desired result. □

We provide the expansion of $L(\lambda, \mu, z)$ without proof as it can be derived in a similar manner. We note that the full proof of the following lemma is provided here [166, Theorem 2, p. 70].

Theorem A.11. *Let* $0 < c < 1$ *and let* λ *and* μ *be any real numbers with* $\lambda > 0$*. We have*

$$L(\lambda, \mu, z) \sim z^{-\lambda}(\log(z))^{\mu} \sum_{r=0}^{\infty} (-1)^r \binom{\mu}{r} \Gamma^{(r)}(\lambda)(\log(z))^{-r}$$

as $z \to \infty$*, where* $\Gamma^{(r)}$ *denotes the* r*th derivative of the gamma function.*

Proof. Very similar to the proof of Theorem A.10. □

Appendix B

Tables and Figures

B.1 Problem set

Table B.1 contains the explicit formula, domain and global minimum of the functions used to generate the high-dimensional test set. The problem set contains 19 problems taken from [62, 49, 147]. Problems that cannot be solved by BARON are marked with '*'. Problems that will not be solved by KNITRO are marked with '°'.

Table B.1: The problem set listed in alphabetical order.

Function	Domain	Global minima
1) Beale [49]	$\boldsymbol{x} \in [-4.5, 4.5]^2$	$g(\boldsymbol{x}^*) = 0$
2) *Branin [49]	$x_1 \in [-5, 10]$ $x_2 \in [0, 15]$	$g(\boldsymbol{x}^*) = 0.397887$
3) Brent [62]	$\boldsymbol{x} \in [-10, 10]^2$	$g(\boldsymbol{x}^*) = 0$
4) °Bukin N.6 [147]	$x_1 \in [-15, -5]$ $x_2 \in [-3, 3]$	$g(\boldsymbol{x}^*) = 0$
5) *Easom [49]	$\boldsymbol{x} \in [-100, 100]^2$	$g(\boldsymbol{x}^*) = -1$
6) Goldstein-Price [49]	$\boldsymbol{x} \in [-2, 2]^2$	$g(\boldsymbol{x}^*) = 3$
7) Hartmann 3 [49]	$\boldsymbol{x} \in [0, 1]^3$	$g(\boldsymbol{x}^*) = -3.86278$
8) Hartmann 6 [49]	$\boldsymbol{x} \in [0, 1]^6$	$g(\boldsymbol{x}^*) = -3.32237$
9) *Levy [147]	$\boldsymbol{x} \in [-10, 10]^4$	$g(\boldsymbol{x}^*) = 0$
10) Perm 4, 0.5 [147]	$\boldsymbol{x} \in [-4, 4]^4$	$g(\boldsymbol{x}^*) = 0$
11) Rosenbrock [147]	$\boldsymbol{x} \in [-5, 10]^3$	$g(\boldsymbol{x}^*) = 0$
12) Shekel 5 [147]	$\boldsymbol{x} \in [0, 10]^4$	$g(\boldsymbol{x}^*) = -10.1532$
13) Shekel 7 [147]	$\boldsymbol{x} \in [0, 10]^4$	$g(\boldsymbol{x}^*) = -10.4029$
14) Shekel 10 [147]	$\boldsymbol{x} \in [0, 10]^4$	$g(\boldsymbol{x}^*) = -10.5364$
15) *Shubert [147]	$\boldsymbol{x} \in [-10, 10]^2$	$g(\boldsymbol{x}^*) = -186.7309$
16) Six-hump camel [147]	$x_1 \in [-3, 3]$ $x_2 \in [-2, 2]$	$g(\boldsymbol{x}^*) = -1.0316$
17) Styblinski-Tang [147]	$\boldsymbol{x} \in [-5, 5]^4$	$g(\boldsymbol{x}^*) = -156.664$
18) Trid [147]	$\boldsymbol{x} \in [-25, 25]^5$	$g(\boldsymbol{x}^*) = -30$
19) Zettl [49]	$\boldsymbol{x} \in [-5, 5]^2$	$g(\boldsymbol{x}^*) = -0.00379$

B.2 Additional experiments

We conducted three more experiments to test REGO's robustness to changes in the parameters. In all three experiments, the same budget and termination criteria as in the main experiment are used.

(A) In this experiment, we assume that no good estimate for μ is known and that μ can be as large as $\sqrt{D}$ (for example, when $\mathcal{X} = [-1, 1]^D$ constraint is imposed). We test REGO with the following parameters: $(d_e, 8.0 \times \sqrt{D}), (d_e + 1, 2.2 \times \sqrt{D}), (d_e + 2, 1.3 \times \sqrt{D})$ and $(d_e + 3, 1.0 \times \sqrt{D})$. Results are presented in Figure B.1 in Appendix B.2.

(B) We fix δ to be $7.5\sqrt{d_e}$ and vary d. The following parameters are used: $(d_e, 7.5 \times \sqrt{d_e}), (d_e + 1, 7.5 \times \sqrt{d_e}), (d_e + 2, 7.5 \times \sqrt{d_e})$ and $(d_e + 3, 7.5 \times \sqrt{d_e})$. Results are presented in Figure B.2 in Appendix B.2.

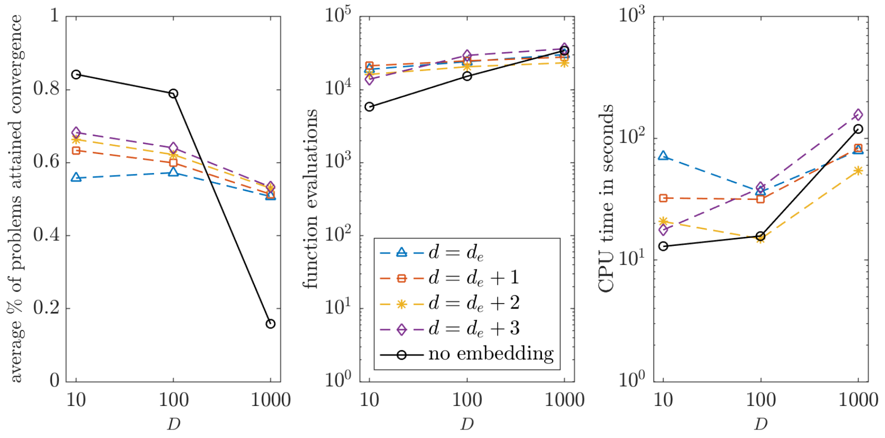
(C) We fix $d = d_e + 1$ and vary δ $(= 5\sqrt{d_e}, 7.5\sqrt{d_e}, 10\sqrt{d_e})$. The following parameters are used: $(d_e + 1, 5 \times \sqrt{d_e}), (d_e + 1, 7.5 \times \sqrt{d_e})$ and $(d_e + 1, 10 \times \sqrt{d_e})$. In the figures we also include curves for $\delta_{opt} = 2.2\sqrt{d_e}$ taken from the main experiment. Results are presented in Figure B.3 in Appendix B.2.

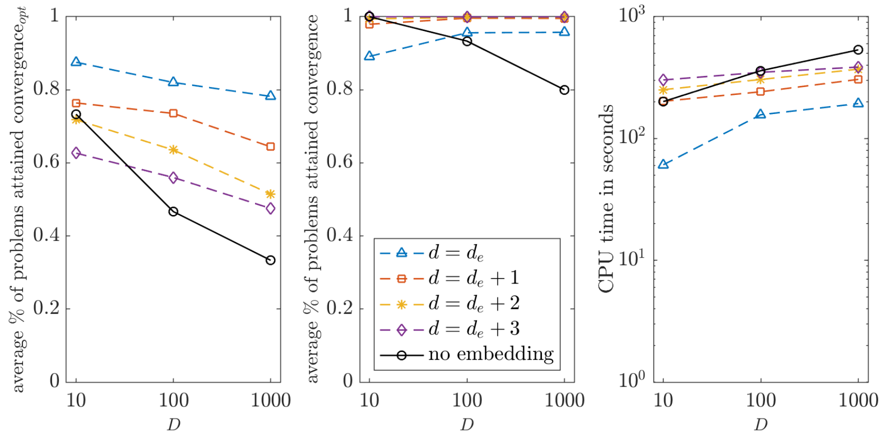
Conclusions

(A) We test robustness of REGO assuming that μ is equal to $\sqrt{D}$ (which makes δ to be relatively large and dependent on D). Despite this dependence, the frequency of convergence for BARON and KNITRO is high showing mild dependence on D.

(B) The purpose of this experiment is to see how different values of d affect the performance of REGO while δ is kept constant. For larger d, we expect (RP) to be successful with higher chance. Nonetheless, the results show that sometimes, for larger d, REGO's performance may be compromised; this is for example true for BARON's convergence$_{opt}$. Since δ is set to a relatively large value, (RP) is successful with high probability even for smallest d. This suggests that as long as d and δ produce relatively high chance of success of (RP), one should stop increasing their values lest convergence to the global minimum require larger computational resources.

(C) In this experiment, we apply REGO with different values of δ while keeping d constant. The results display no significant differences between the performances with different parameters. Even the results with the optimal δ (used in the main experiment) do not differ considerably from the one with the largest δ except for BARON where the former wins in terms of convergence$_{opt}$ and CPU time. The results of this experiment together with the results in (B) indicate that it is better to increase δ and keep d constant if one wants to increase success of (RP) with minimal increase in computational cost.


a) DIRECT


b) BARON

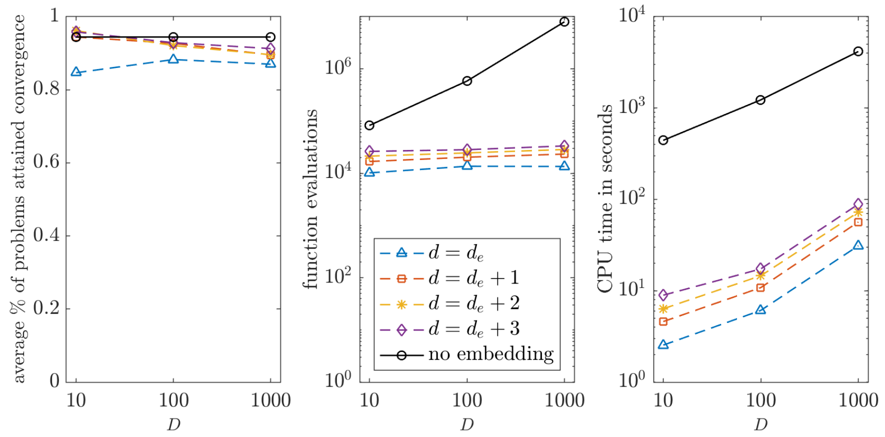
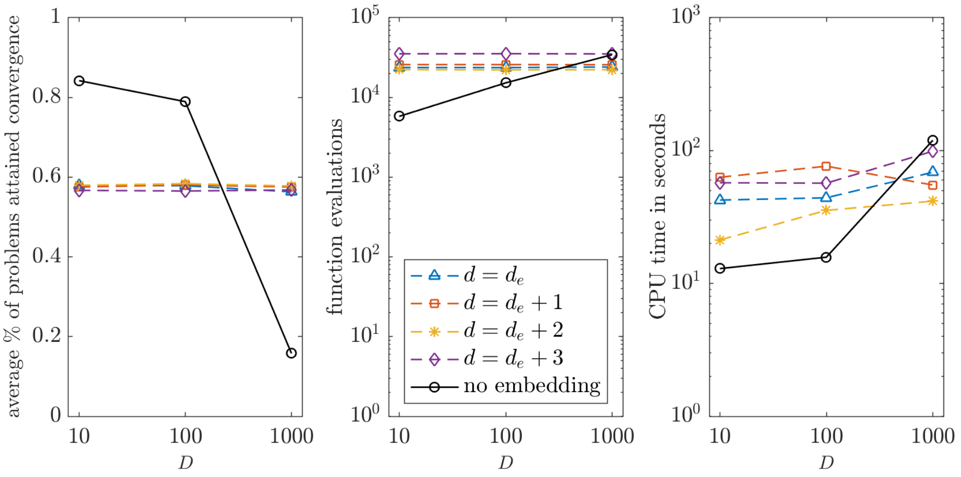
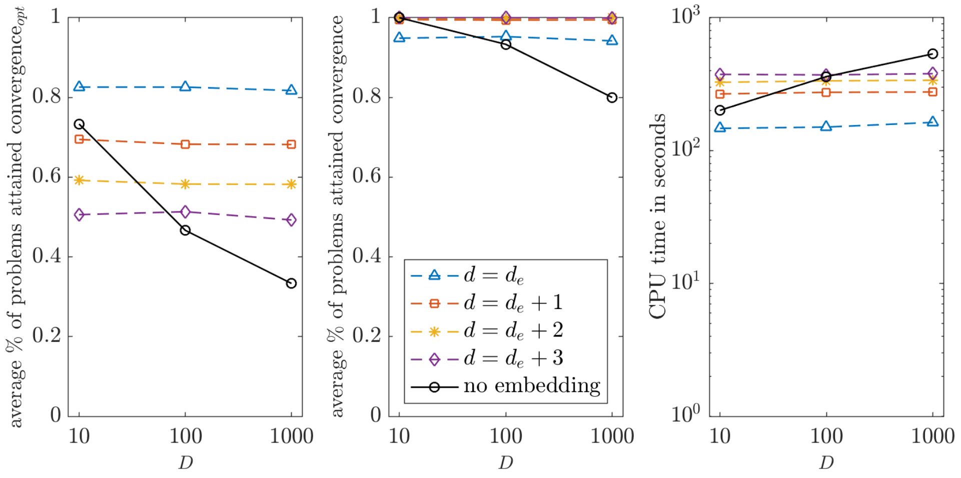

c) KNITRO

Figure B.1: Experiment A: REGO with DIRECT, BARON and KNITRO with $(d, \delta) = (d_e, 8.0 \times \sqrt{D}), (d_e + 1, 2.2 \times \sqrt{D}), (d_e + 2, 1.3 \times \sqrt{D})$ and $(d_e + 3, 1.0 \times \sqrt{D})$.

a) DIRECT



b) BARON



c) KNITRO

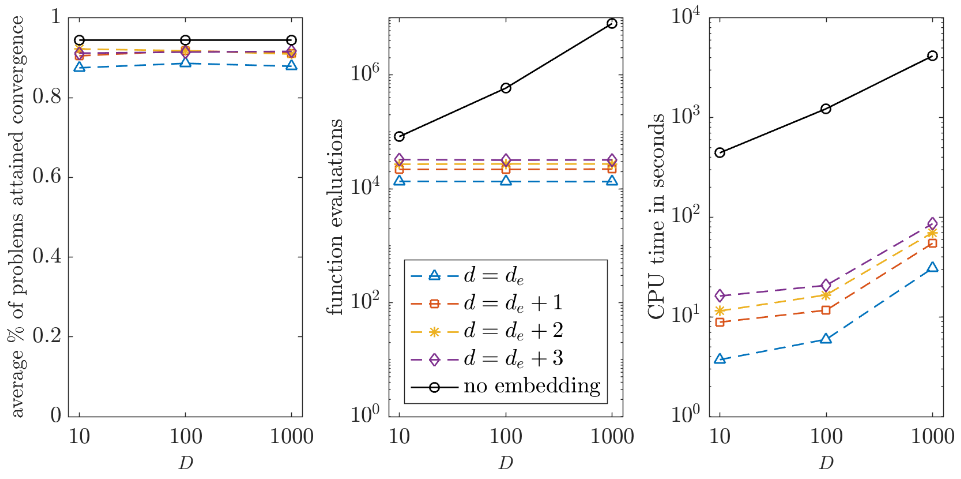
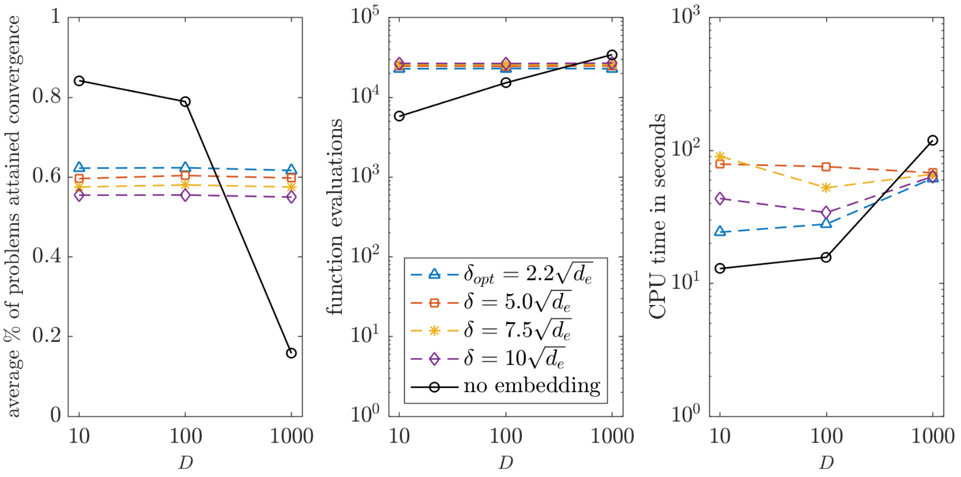
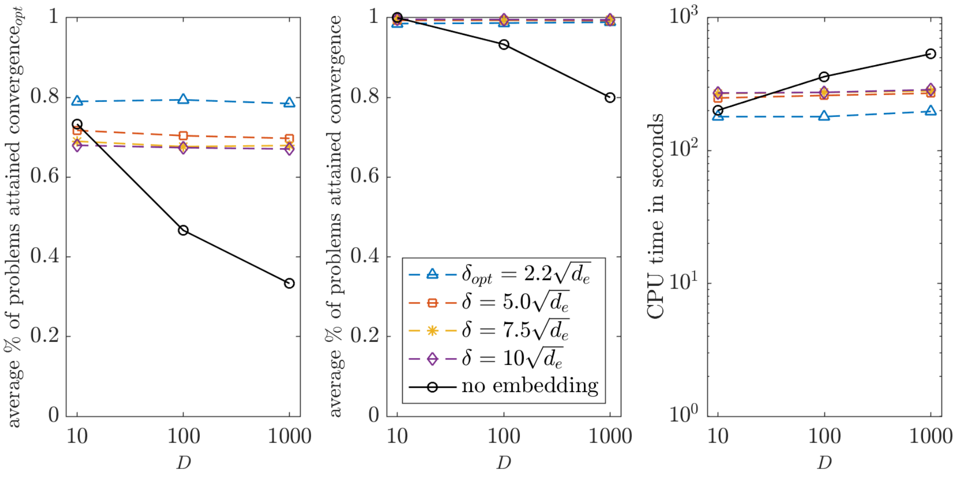


Figure B.2: Experiment B: REGO with DIRECT, BARON and KNITRO with $\delta = 7.5\sqrt{d_e}$ fixed and $d = d_e$, $d_e + 1$, $d_e + 2$ and $d_e + 3$.

a) DIRECT



b) BARON



c) KNITRO

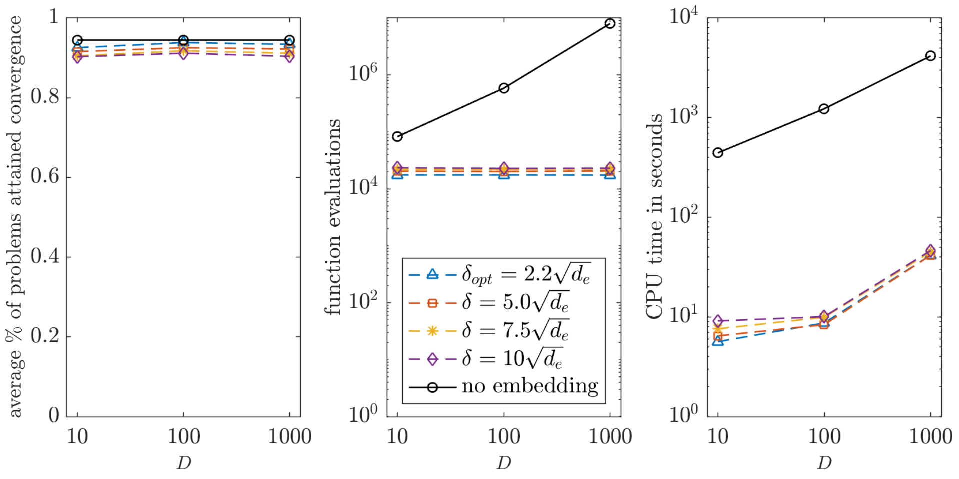


Figure B.3: Experiment C: REGO with DIRECT, BARON and KNITRO with $d = d_e + 1$ fixed and $\delta = 5\sqrt{d_e}, 7.5\sqrt{d_e}, 10\sqrt{d_e}$ and $2.2\sqrt{d_e}$ (δ_{opt}).