Constrained global optimization of functions with low effective dimensionality using multiple random embeddings

Coralia Cartis [*, ‡] Estelle Massart [*, §] Adilet Otemissov [*, ‡]

1st October 2020

Abstract

We consider the bound-constrained global optimization of functions with low effective dimensionality, that are constant along an (unknown) linear subspace and only vary over the effective (complement) subspace. We aim to implicitly explore the intrinsic low dimensionality of the constrained landscape using feasible random embeddings, in order to understand and improve the scalability of algorithms for the global optimization of these special-structure problems. A reduced subproblem formulation is investigated that solves the original problem over a 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 an ϵ-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.

Keywords: global optimization, constrained optimization, random embeddings, dimensionality reduction techniques, functions with low effective dimensionality.

1 Introduction

In this paper, we address the bound-constrained global optimization problem

$$f^* := \min_{\boldsymbol{x} \in \mathcal{X}} \; f(\boldsymbol{x}), \tag{P}$$

[*]The order of the authors is alphabetical; the third author is the primary contributor. Mathematical Institute, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG, UK; cartis,massart,otemissov@maths.ox.ac.uk

[‡]The Alan Turing Institute, The British Library, London, NW1 2DB, UK. This work was supported by The Alan Turing Institute under The Engineering and Physical Sciences Research Council (EPSRC) grant EP/N510129/1 and under the Turing project scheme.

[§]National Physical Laboratory, Hampton Road, Teddington, Middlesex, TW11 0LW, UK. This author's work was supported by the National Physical Laboratory.

where $f : \mathbb{R}^D \to \mathbb{R}$ is continuous, possibly non-convex and deterministic[1], and where, without loss of generality, $\mathcal{X} := [-1, 1]^D \subseteq \mathbb{R}^D$.

In an attempt to alleviate the curse of dimensionality of generic global optimization, we focus on objective functions with '*low effective dimensionality*' [54], namely, those that only vary over a low-dimensional 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*' [12] or '*multi-ridge*' [23, 52]. They are frequently encountered in applications, typically when tuning (over)parametrized models and processes, such as in hyper-parameter optimization for neural networks [3], heuristic algorithms for combinatorial optimization problems [32], complex engineering and physical simulation problems [12] as in climate modelling [35], and policy search and dynamical system control [57, 24].

When the objective has low effective dimensionality and the effective subspace of variation is known, it is straightforward to cast (P) into a lower-dimensional problem which has the same global minimum f^* by restricting it to and solving (P) only within this important subspace. Typically, however, the effective subspace is unknown, and random embeddings have been proposed to reduce the size of (P) and hence the cost of its solution, while attempting to preserve the problem's (original) global minimum values. In this paper, we investigate the following feasible formulation of the reduced randomised problem,

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

where $\boldsymbol{A}$ is a $D \times d$ Gaussian random matrix (see Definition A.1) with $d \ll D$, and where $\boldsymbol{p} \in \mathcal{X}$ is user-defined and provides additional flexibility that we exploit algorithmically. Our approach needs the following clarification.

Definition 1.1. 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 $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} \in \mathcal{X}$.

We derive a lower bound on the probability that (RP$\mathcal{X}$) is successful in the case when d is equal to or larger than the effective dimension. We show that this success probability is positive and that it depends on both the effective subspace and the ambient dimensions[2]. However, in the case when the effective subspace is aligned with the coordinate axes, we show that the dependence on D in this lower bound is at worst algebraic. We then propose X-REGO ($\mathcal{X}$ - Random Embeddings for Global Optimization), a generic algorithmic framework for solving (P) using multiple random embeddings. Namely, X-REGO solves (RP$\mathcal{X}$) repeatedly with different $\boldsymbol{A}$ and possibly different $\boldsymbol{p}$, and can use any global optimization algorithm for solving the reduced problem (RP$\mathcal{X}$). Using the computed lower bound on the probability of success of (RP$\mathcal{X}$), we derive a global convergence result for X-REGO, showing that as the number of random embeddings increases, X-REGO converges linearly, with probability one, to an ϵ-neighbourhood of a global minimizer of (P).

Existing relevant literature. Optimization of functions with low effective dimensionality has been recently studied primarily as an attempt to remedy the scalability challenges of Bayesian Optimization (BO), such as in [15, 54, 26, 39, 19]. Investigations of these special-structure

[1] Our analysis would be significantly more involved, but still possible, if f is only well defined on $\mathcal{X}$. Note that in our X-REGO algorithm, we only query $f(\boldsymbol{x})$ at feasible points $\boldsymbol{x} \in \mathcal{X}$.

[2] A brief description, without proofs, of a subset of the main results of this paper has appeared as (a sub)part of a four-page conference proceedings paper (without any supplementary materials) in the ICML Workshop "Beyond first order methods in ML systems" (2020), see https://drive.google.com/file/d/1JxQc9rSK8GYchKnDp0dhwEa4f3AeyNeb/view.

problems have been extended beyond BO, to derivative-free optimization [45], multi-objective optimization [44] and evolutionary methods [47, 13]. As the effective subspace is generally unknown, some existing approaches learn the effective subspace beforehand [23, 52, 15, 19], while others estimate it during the optimization, updating the estimate as new information becomes available on the objective function [26, 57, 11, 13]. We focus here on an alternative approach, bypassing the subspace learning phase, and optimizing directly over random low-dimensional subspaces, as proposed in [54, 6, 7, 34].

Wang et al. [54] propose the REMBO algorithm, that solves, using Bayesian methods, a single reduced subproblem,

$$\begin{aligned} &\min_{\boldsymbol{y}} \ f(\boldsymbol{A}\boldsymbol{y}) \\ &\text{subject to } \ \boldsymbol{y} \in \mathcal{Y} = [-\delta, \delta]^d, \end{aligned} \tag{RP}$$

where $\boldsymbol{A}$ is as above, and $\delta > 0$. They evaluate the probability that the solution of (RP) corresponds to a solution of the original problem (P) in the case when the effective subspace is aligned with coordinate axes and when $d = d_e$, where d_e denotes the dimension of the effective subspace; they show that this probability of success of (RP) depends on the parameter δ (the size of the $\mathcal{Y}$ box), and it decreases as δ shrinks. Conversely, setting δ large may result in large computational costs to solve (RP). Thus, a careful calibration of δ is needed for good algorithmic performance. The theoretical analysis in [54] has been extended by Sanyang and Kabán [47], where the probability of success of (RP) is quantified in the case $d \geq d_e$; an algorithm, called REMEDA, is also proposed in [47] that uses Gaussian random embeddings in the framework of evolutionary methods for high-dimensional unconstrained global optimization.

In the recent preprint [9], we further extend these analyses to arbitrary effective subspaces (i.e., not necessarily aligned with the coordinate axes) and random embeddings of dimension $d \geq d_e$, and consider the wider framework of generic unconstrained high-dimensional global optimization. We propose the REGO algorithm, that replaces the high-dimensional problem (P) (with $\mathcal{X} = \mathbb{R}^D$), by a *single* reduced problem (RP), and solves (RP) using any global optimization algorithm. Instead of estimating solely the norm of an optimal solution of (RP), as in [54, 47], we derive its exact probability distribution. Furthermore, we show that its squared Euclidean norm (when appropriately scaled) follows an inverse chi-squared distribution with $d - d_e + 1$ degrees of freedom, and use a tail bound on the chi-squared distribution to get a lower bound on the probability of success of (RP). Our theory and numerical experiments indicate that, under suitable assumptions, the success of (RP) is essentially independent on D, but depends mainly on two factors: the gap between the subspace dimension d and the effective dimension d_e, and the ratio between δ (the size of the low-dimensional domain), and the distance from the origin (the centre of the original domain $\mathcal{X}$) to the closest affine subspace of global minimizers.

In contrast to [47] and [9], the present case of the *constrained* problem (P) poses a new challenge: a solution $\boldsymbol{y}^*$ of (RP) is not necessarily feasible for the full-dimensional problem (P) (i.e., $\boldsymbol{A}\boldsymbol{y}^* \notin \mathcal{X}$). To remedy this, Wang et al. [54] 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) 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 [6, 7]. Binois et al. [7] further advances the discussion regarding the choice of the low-dimensional domain $\mathcal{Y}$ in (RP) and computes an 'optimal' set $\mathcal{Y}^* \subset \mathbb{R}^d$, i.e., a set that has minimum (here, infimum) volume among all the sets $\mathcal{Y} \subset \mathbb{R}^d$ for which the image of the mapping $\mathcal{Y} \to \mathcal{X} : \boldsymbol{y} \mapsto p_{\mathcal{X}}(\boldsymbol{A}\boldsymbol{y})$ contains the 'maximal embedded set' $\{p_{\mathcal{X}}(\boldsymbol{A}\boldsymbol{y}) : \boldsymbol{y} \in \mathbb{R}^d\}$, where $p_{\mathcal{X}}(\boldsymbol{x})$ is the classical Euclidean projection of $\boldsymbol{x}$ on $\mathcal{X}$. They show that $\mathcal{Y}^*$ has an intricate representation when the dimension of the full-dimensional problem is large, and propose to replace the Euclidean projection map $p_{\mathcal{X}}$ suggested by Wang et al. [54] by an alternative mapping for which an 'optimal' low-dimensional domain has nicer properties. Nayebi

et al. [40] 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 need to combine optimization algorithms that rely on random Gaussian embeddings with a projection step has also been recently discussed in [38], where it is suggested to replace the formulation (RP) by (RP$\mathcal{X}$), that we also consider in this paper. However, Letham et al. [38] do not provide analytical estimates of the probability of success of this new formulation, solely evaluating it numerically using Monte-Carlo simulations; they also do not use multiple random embeddings. Our proposed X-REGO algorithmic framework (and more precisely, the adaptive variant A-REGO described in Section 5) is closely related to the sequential algorithm proposed by Qian et al. [45], in the framework of unconstrained derivative-free optimization of functions with approximate low-effective dimensionality, and to the algorithm proposed in [34] for constrained Bayesian optimization of functions with low-effective dimension, using one-dimensional random embeddings. However, our results rely on the assumption that the subspace dimension d is larger than the effective dimension d_e, and so our approach significantly differs from [34]. Very recently, Tran-The et al. [51] have proposed an algorithm that uses several low-dimensional (deterministic) embeddings in parallel for Bayesian optimization of high-dimensional functions.

Randomized subspace methods have recently attracted much interest for local or convex optimization problems; see for example, [41, 36, 28, 31]; no low effective dimensionality assumption is made in these works. Finally, we note that the main step in our convergence analysis consists in deriving a lower bound on the probability that a random subspace of given dimension intersects a given set (the set of approximate global minimizers), which is an important problem in stochastic geometry, see, e.g., the extensive discussion by Oymak and Tropp [43]. Unlike the results presented in [43], our results do not involve statistical dimensions of sets, which are unknown and, in our case, problem dependent.

Our contributions. Here we investigate a general random embedding framework for the *bound-constrained* global optimization of functions with low effective dimensionality. This framework replaces the original, potentially high-dimensional problem (P) with several reduced and randomized subproblems of the form (RP$\mathcal{X}$), which directly ensures feasibility of the iterates with respect to the constraints.

Using various properties of Gaussian matrices and a useful result from [9], we 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 $\boldsymbol{A}$, the parameter $\boldsymbol{p}$ of (RP$\mathcal{X}$) and an arbitrary global minimizer $\boldsymbol{x}^*$ of (P). We show that $\boldsymbol{w}$ follows a $(D - d_e)-$dimensional t-distribution with $d - d_e + 1$ degrees of freedom, and provide a lower bound on the probability of success of (RP$\mathcal{X}$) in terms of the integral of the probability density function of $\boldsymbol{w}$ over a given closed domain. In the case when the effective subspace is aligned with the coordinate axes, the closed domain simplifies to a $(D - d_e)-$dimensional box, and we provide an asymptotic expansion of the integral of the probability density function over the box, 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 $\boldsymbol{p}$.

We also propose the X-REGO algorithm, a generic framework for the constrained global optimization problem (P) that sequentially or in parallel solves multiple subproblems (RP$\mathcal{X}$), varying $\boldsymbol{A}$ and also possibly $\boldsymbol{p}$. We prove global convergence of X-REGO to a set of approx-

imate 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 used to solve the reduced problem (namely, it must solve (RP$\mathcal{X}$) globally and approximately, to required accuracy), and allows a diverse set of possible choices of $\boldsymbol{p}$ (random, fixed, adaptive, deterministic). Our convergence proof crucially uses our result that the probability of success of (RP$\mathcal{X}$) is positive and uniformly bounded away from zero with respect to the choice of $\boldsymbol{p}$, and hence, assumes that $d \geq d_e$.

We provide an extensive numerical comparison of several variants of X-REGO on a set of test problems with low effective dimensionality, using three different solvers for (RP$\mathcal{X}$), namely, BARON [46], DIRECT [22] and (global and local) KNITRO [8]. 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.

Paper outline. In Section 2, we recall the definition of functions with low effective dimensionality and some existing results that we will use in our analysis. Section 3 derives lower bounds for the probability of success of (RP$\mathcal{X}$). The X-REGO algorithm and its global convergence are then presented in Section 4, while in Section 5, different X-REGO variants are compared numerically on benchmark problems using three optimization solvers (DIRECT, BARON and KNITRO) for the subproblems. Our conclusions are drawn in Section 6.

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 $\langle \cdot, \cdot \rangle$, $\| \cdot \|$ and $\| \cdot \|_\infty$ for the usual Euclidean inner product, 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 A.1) and write χ^2_n to denote a chi-squared random variable with n degrees of freedom (see Definition A.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 $x \geq x_0$, $M_1 g(x) \leq |f(x)| \leq M_2 g(x)$.

2 Preliminaries

2.1 Functions with low effective dimensionality

Definition 2.1 (Functions with low effective dimensionality [54])**.** 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

$$f(\boldsymbol{x}_\top + \boldsymbol{x}_\perp) = f(\boldsymbol{x}_\top), \tag{2.1}$$

and d_e is the smallest integer satisfying (2.1).

The linear subspaces $\mathcal{T}$ and $\mathcal{T}^\perp$ are called the *effective* and *constant* subspaces of f, respectively. In this paper, we make the following assumption on the function f.

Assumption 2.2. *The function $f : \mathbb{R}^D \to \mathbb{R}$ is continuous and has effective dimensionality d_e such that $d_e < D$ and $d_e \leq d$, with effective subspace[3] $\mathcal{T}$ 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 let $\boldsymbol{x}_\top = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}$ 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.*

We define the set of feasible global minimizers of problem (P),

$$G := \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) = f^*\}. \tag{2.2}$$

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

$$f^* = f(\boldsymbol{x}^*) = f(\boldsymbol{x}^*_\top + \tilde{\boldsymbol{x}}) = f(\boldsymbol{x}^*_\top). \tag{2.3}$$

The minimizer $\boldsymbol{x}^*_\top$ may lie outside $\mathcal{X}$, and furthermore, there may be multiple points $\boldsymbol{x}^*_\top$ in $\mathcal{T}$ satisfying $f^* = f(\boldsymbol{x}^*_\top)$ as illustrated in [9, Example 1.1]. Thus, the set G is (generally)[4] a union of (possibly infinitely many) $(D-d_e)$-dimensional simply-connected polyhedral sets, each corresponding to a particular $\boldsymbol{x}^*_\top$. If $\boldsymbol{x}^*_\top$ is unique, i.e., every global minimizer $\boldsymbol{x}^* \in G$ has the same Euclidean projection $\boldsymbol{x}^*_\top$ on the effective subspace, then G is the $(D-d_e)$-dimensional set $\{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{x} \in \boldsymbol{x}^*_\top + \mathcal{T}^\perp\}$.

Definition 2.3. Suppose Assumption 2.2 holds. For any global minimizer $\boldsymbol{x}^* \in G$, let $G^* := \{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{x} \in \boldsymbol{x}^*_\top + \mathcal{T}^\perp\}$ be the simply connected subset of G that contains $\boldsymbol{x}^*_\top = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}^*$, and $\mathcal{G}^* := \{\boldsymbol{x} \in \mathbb{R}^D : \boldsymbol{x} \in \boldsymbol{x}^*_\top + \mathcal{T}^\perp\}$, the $(D-d_e)$-dimensional affine subspace that contains G^*.

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 $\boldsymbol{V}$ is defined in Assumption 2.2. For each G^*, we define the corresponding set of "admissible" $(D-d_e)$-dimensional vectors as

$$\bar{G}^* := \{\boldsymbol{g} \in \mathbb{R}^{D-d_e} : \boldsymbol{x}^*_\top + \boldsymbol{V}\boldsymbol{g} \in G^*\}. \tag{2.4}$$

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 2.3 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, $\mathrm{Vol}(\bar{G}^*) = 0$; a case we need to sometimes exclude from our analysis.

Definition 2.4. Let G^* and $\bar{G}^*$ be defined as in Definition 2.3 and (2.4), respectively. We say that G^* is non-degenerate if $\mathrm{Vol}(\bar{G}^*) > 0$.

The definitions and assumptions introduced in this section are illustrated next in Figure 1.

Geometric description of the problem. Figure 1 sketches the linear mapping $\boldsymbol{y} \to \boldsymbol{A}\boldsymbol{y} + \boldsymbol{p}$ that maps points from $\mathbb{R}^d$ to points in the affine subspace $\boldsymbol{p} + \mathrm{range}(\boldsymbol{A})$ in $\mathbb{R}^D$. This figure also illustrates the case of a non-degenerate simply-connected component G^* of global minimizers (blue line; Definition 2.3), which here has dimension $D - d_e = 1$. Degeneracy of G^* (Definition 2.4) would occur if $\boldsymbol{x}^*_\top$ was a vertex of the domain $\mathcal{X}$, in which case the corresponding G^* would be a singleton.

[3]Note that $\mathcal{T}$ in Assumption 2.2 may not be aligned with the standard axes.

[4]Except in degenerate cases, see Definition 2.4.

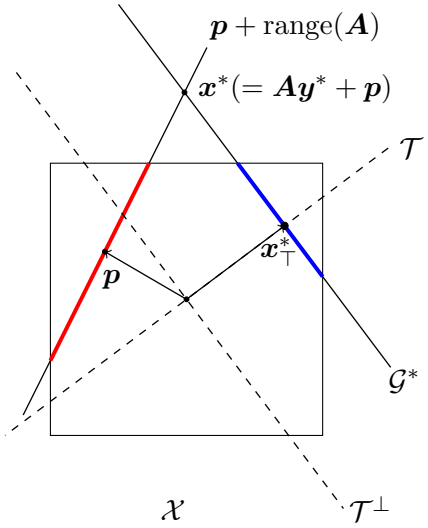

Figure 1: Abstract illustration of the embedding of an affine d-dimensional subspace $\boldsymbol{p}+\text{range}(\boldsymbol{A})$ into $\mathbb{R}^D$. The red line represents the feasible set of solutions along $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ and the blue line represents the set G^*. The random subspace intersects $\mathcal{G}^*$ at $\boldsymbol{x}^*$, which is infeasible.

For (RP$\mathcal{X}$) to be successful in solving the original problem (P), Figure 1 illustrates it is sufficient that the red line segment (the feasible set of (reduced) solutions in $\mathbb{R}^d$ mapped to $\mathbb{R}^D$) intersects the blue line segment (the set G^*)[5]. The blue and red line segments do not intersect in Figure 1, but their prolongations outside $\mathcal{X}$ ($\mathcal{G}^*$ and $\boldsymbol{p} + \text{range}(\boldsymbol{A})$) do[6]. In Section 2.2, we review an existing characterization for a reduced minimizer ($\boldsymbol{y}^*$ in Figure 1), thus quantifying a specific intersection between the random subspace and $\mathcal{G}^*$. We then use this characterization in Section 3 to derive a lower bound on the probability of $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}$ to belong to G^*, namely, to be feasible for the original problem (P).

2.2 Characterization of (unconstrained) minimizers in the reduced space

This section summarizes results from [9] that characterize the distribution of a random reduced minimizer $\boldsymbol{y}^*$ such that $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}$ is an unconstrained minimizer of f.

Let $\mathcal{S}^* := \{\boldsymbol{y}^* \in \mathbb{R}^d : \boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} \in \mathcal{G}^*\}$, with $\mathcal{G}^*$ defined in Definition 2.3, be a subset of points $\boldsymbol{y}^*$ corresponding to solutions of minimizing f over the entire $\mathbb{R}^D$. With probability one, $\mathcal{S}^*$ is a singleton if $d = d_e$ and has infinitely many points if $d > d_e$ [9, Corollary 3.3]. It is sufficient to find one of the reduced minimizers in $\mathcal{S}^*$, ideally one that is easy to analyse, and that is close to the origin (i.e., the centre of the domain $\mathcal{X}$) in some norm so as to encourage the feasibility with respect to $\mathcal{X}$ of its image through $\boldsymbol{A}$. An obvious candidate is the minimal Euclidean norm solution,

$$\begin{aligned} \boldsymbol{y}_2^* = \underset{\boldsymbol{y}\in\mathbb{R}^d}{\text{argmin}} \;\; & \|\boldsymbol{y}\|_2 \\ \text{s.t.} \;\; & \boldsymbol{y} \in \mathcal{S}^*. \end{aligned} \tag{2.5}$$

Theorem 2.5. *[9, Theorem 3.1] Suppose Assumption 2.2 holds. Let $\boldsymbol{x}^*$ be any global minimizer of* (P) *with Euclidean projection $\boldsymbol{x}^*_\top$ on the effective subspace, and $\boldsymbol{p} \in \mathcal{X}$, a given vector. Let*

[5] If $G^* = G$, this sufficient condition is also necessary; else, we need to check the other simply connected components of G to decide whether (RP$\mathcal{X}$) is successful or not.

[6] This is related to [54, Theorem 2], which says that if the dimension of the embedded subspace (d) is greater than the effective dimension (d_e) of f then $\mathcal{G}^*$ and $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ intersect with probability one. Wang et al. [54] have shown this result for the case $\boldsymbol{p} = \boldsymbol{0}$, but it can easily be generalized to arbitrary $\boldsymbol{p}$.

$\boldsymbol{A}$ *be a* $D \times d$ *Gaussian matrix. Then* $\boldsymbol{y}_2^*$ *defined in* (2.5) *is given by*

$$\boldsymbol{y}_2^* := \boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{z}^*, \tag{2.6}$$

which is the minimum Euclidean norm solution to the system

$$\boldsymbol{B}\boldsymbol{y}^* = \boldsymbol{z}^*, \tag{2.7}$$

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

$$\boldsymbol{U}\boldsymbol{z}^* = \boldsymbol{x}_\top^* - \boldsymbol{p}_\top, \text{ with } \boldsymbol{p}_\top = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{p}. \tag{2.8}$$

Proof. See Appendix B. □

Remark 2.6. Note that $\boldsymbol{B} = \boldsymbol{U}^T\boldsymbol{A}$ is a $d_e \times d$ Gaussian matrix, since $\boldsymbol{U}$ has orthonormal columns (see Theorem A.2). Also, (2.8) implies $\|\boldsymbol{z}^*\|_2 = \|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|_2$.

Using (2.6) and various properties of Gaussian matrices, [9] shows that the squared Euclidean norm of $\boldsymbol{y}_2^*$ follows the (appropriately scaled) inverse chi-squared distribution.

Theorem 2.7. *([9, Theorem 3.7]) Suppose Assumption 2.2 holds. Let* $\boldsymbol{x}^*$ *be any global minimizer of* (P) *and* $\boldsymbol{p} \in \mathcal{X}$ *a given vector, with respective projections* $\boldsymbol{x}_\top^*$ *and* $\boldsymbol{p}_\top$ *on the effective subspace. Let* $\boldsymbol{A}$ *be a* $D \times d$ *Gaussian matrix. Then,* $\boldsymbol{y}_2^*$ *defined in* (2.5) *satisfies*

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

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

3 Estimating the success of the reduced problem

This section derives lower bounds on the probability of success of (RP$\mathcal{X}$). Lemma 3.1 lower bounds this probability by that of a non-empty intersection between the random subspace $\boldsymbol{p}+$ range($\boldsymbol{A}$) and an arbitrary simply-connected component G^* of the set of global minimizers (Definition 2.3). This probability is further expressed in Corollary 3.4 in terms of a random vector $\boldsymbol{w}$ that follows a multivariate t-distribution. From Section 3.1 onwards, we derive positive and/or quantifiable lower bounds on the probability of success of (RP$\mathcal{X}$), while also trying to eliminate, wherever possible, the dependency of the lower bounds on the choice of $\boldsymbol{p}$ and G^*.

Lemma 3.1. *Suppose Assumption 2.2 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. Let* $\boldsymbol{y}_2^*$ *be defined in* (2.5). *The reduced problem* (RP$\mathcal{X}$) *is successful in the sense of Definition 1.1 if* $\boldsymbol{A}\boldsymbol{y}_2^* + \boldsymbol{p} \in \mathcal{X}$, *namely*

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is successful}] \geq \mathbb{P}[-\boldsymbol{1} \leq \boldsymbol{A}\boldsymbol{y}_2^* + \boldsymbol{p} \leq \boldsymbol{1}]. \tag{3.1}$$

Proof. This is an immediate consequence of Definition 1.1 and (2.5), as the latter implies $\boldsymbol{A}\boldsymbol{y}_2^* + \boldsymbol{p} \in \mathcal{G}^*$ and so $f(\boldsymbol{A}\boldsymbol{y}_2^* + \boldsymbol{p}) = f^*$. □

Let us further express (3.1) as follows. Let $\boldsymbol{Q} = (\boldsymbol{U}\ \boldsymbol{V})$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are defined in Assumption 2.2. Since $\boldsymbol{Q}$ is orthogonal, we have

$$\boldsymbol{A}\boldsymbol{y}_2^* = \boldsymbol{Q}\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{y}_2^* = \boldsymbol{Q}\begin{pmatrix}\boldsymbol{U}^T \\ \boldsymbol{V}^T\end{pmatrix}\boldsymbol{A}\boldsymbol{y}_2^*. \tag{3.2}$$

Using (2.6), we get $\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{y}_2^* = \boldsymbol{z}^*$. Letting

$$\boldsymbol{w} := \boldsymbol{V}^T\boldsymbol{A}\boldsymbol{y}_2^*, \tag{3.3}$$

we get

$$\boldsymbol{A}\boldsymbol{y}_2^* = \boldsymbol{Q}\begin{pmatrix}\boldsymbol{z}^*\\ \boldsymbol{w}\end{pmatrix} = \begin{pmatrix}\boldsymbol{U} & \boldsymbol{V}\end{pmatrix}\begin{pmatrix}\boldsymbol{z}^*\\ \boldsymbol{w}\end{pmatrix} = \boldsymbol{U}\boldsymbol{z}^* + \boldsymbol{V}\boldsymbol{w} = \boldsymbol{x}_\top^* - \boldsymbol{p}_\top + \boldsymbol{V}\boldsymbol{w}, \tag{3.4}$$

where in the last equality, we used (2.8). By substituting $\boldsymbol{p} = \boldsymbol{p}_\top + \boldsymbol{p}_\perp$ and (3.4) in (3.1), we obtain

$$\begin{aligned}\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] &\geq \mathbb{P}[-\mathbf{1} \leq \boldsymbol{A}\boldsymbol{y}_2^* + \boldsymbol{p} \leq \mathbf{1}]\\ &= \mathbb{P}[-\mathbf{1} \leq \boldsymbol{x}_\top^* - \boldsymbol{p}_\top + \boldsymbol{V}\boldsymbol{w} + \boldsymbol{p}_\top + \boldsymbol{p}_\perp \leq \mathbf{1}]\\ &= \mathbb{P}[-\mathbf{1} \leq \boldsymbol{x}_\top^* + \boldsymbol{p}_\perp + \boldsymbol{V}\boldsymbol{w} \leq \mathbf{1}].\end{aligned} \tag{3.5}$$

According to this derivation, all the randomness within the lower bound (3.5) is contained in the random vector $\boldsymbol{w}$. The next theorem, derived in Appendix C, provides the probability density function of this random vector.

Remark 3.2. Suppose that Assumption 2.2 holds and recall (2.2). If there exists $\boldsymbol{x}^* \in G$ such that $\boldsymbol{x}_\top^* = \boldsymbol{p}_\top$, where the subscript represents the respective Euclidean projections on the effective subspace, then $f(\boldsymbol{p}) = f(\boldsymbol{p}_\top + \boldsymbol{p}_\perp) = f(\boldsymbol{x}_\top^* + \boldsymbol{p}_\perp) = f^*$, where $\boldsymbol{p}_\perp$ is the Euclidean projection of $\boldsymbol{p}$ on the constant subspace $\mathcal{T}^\perp$ of the objective function. Thus $\boldsymbol{p} \in G$ so that, for any embedding $\boldsymbol{A}$, (RP$\mathcal{X}$) is successful with the trivial solution $\boldsymbol{y}^* = \mathbf{0}$. Therefore, in our next result, without loss of generality, we make the assumption $\boldsymbol{x}_\top^* \neq \boldsymbol{p}_\top$.

Theorem 3.3 (The p.d.f. of $\boldsymbol{w}$)**.** *Suppose that Assumption 2.2 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. Assume that $\boldsymbol{p}_\top \neq \boldsymbol{x}_\top^*$, where the subscript represents the Euclidean projection on the effective subspace. The random vector $\boldsymbol{w}$ defined in* (3.3) *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*

$$g(\bar{\boldsymbol{w}}) = \frac{1}{(\sqrt{\pi}\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|)^m}\left[\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{n}{2})}\right]\left(1 + \frac{\bar{\boldsymbol{w}}^T\bar{\boldsymbol{w}}}{\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\|^2}\right)^{-(m+n)/2}, \tag{3.6}$$

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

Proof. See Appendix C. □

The remainder of this section aims at answering the two following questions: *Is the probability of success of* (RP$\mathcal{X}$) *positive for any $\boldsymbol{p}$? If yes, can we derive a positive lower bound on the probability of success of* (RP$\mathcal{X}$) *that does not depend on $\boldsymbol{p}$?* We show that both questions can be answered positively, and use this extensively in our global convergence analysis in Section 4.

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

We first summarize the above analysis in the following corollary.

Corollary 3.4. *Suppose that Assumption 2.2 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. Assume that $\boldsymbol{p}_\top \neq \boldsymbol{x}_\top^*$, where the subscript represents the Euclidean projection on the effective subspace. Then*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] \geq \mathbb{P}(-\mathbf{1} \leq \boldsymbol{x}_\top^* + \boldsymbol{p}_\perp + \boldsymbol{V}\boldsymbol{w} \leq \mathbf{1}), \tag{3.7}$$

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 (3.1)–(3.5) and Theorem 3.3. □

We need the following additional assumption.

Assumption 3.5. *Assume that Assumption 2.2 holds, and that there is a set G^* defined in Definition 2.3 that is non-degenerate according to Definition 2.4.*

Theorem 3.6. *Suppose that Assumption 3.5 holds, and let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix. Then, for any $\boldsymbol{p} \in \mathcal{X}$,*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] > 0. \tag{3.8}$$

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 $(\mathrm{RP}\mathcal{X})$ yields $f(\boldsymbol{p}) = f^*$.

Assume now that $\boldsymbol{p} \in \mathcal{X} \setminus G$. Assumption 3.5 implies that there exists a global minimizer $\boldsymbol{x}^*$ and associated G^* for which $\mathrm{Vol}(\bar{G}^*) > 0$, where G^* and $\bar{G}^*$ are defined in Definition 2.3 and (2.4), respectively. Using (3.7) with this particular $\boldsymbol{x}^*$ and noting that $\boldsymbol{p}_\perp = \boldsymbol{V}\boldsymbol{V}^T\boldsymbol{p}$ gives us

$$\begin{aligned}
\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] &\geq \mathbb{P}[-\boldsymbol{1} \leq \boldsymbol{x}^*_\top + \boldsymbol{V}(\boldsymbol{V}^T\boldsymbol{p} + \boldsymbol{w}) \leq \boldsymbol{1}] \\
&= \mathbb{P}[\boldsymbol{V}^T\boldsymbol{p} + \boldsymbol{w} \in \{\boldsymbol{g} \in \mathbb{R}^{D-d_e} : -\boldsymbol{1} \leq \boldsymbol{x}^*_\top + \boldsymbol{V}\boldsymbol{g} \leq \boldsymbol{1}\}] \\
&= \mathbb{P}[\boldsymbol{V}^T\boldsymbol{p} + \boldsymbol{w} \in \bar{G}^*] \\
&= \mathbb{P}[\boldsymbol{w} \in -\boldsymbol{V}^T\boldsymbol{p} + \bar{G}^*] \\
&= \int_{-\boldsymbol{V}^T\boldsymbol{p}+\bar{G}^*} g(\bar{\boldsymbol{w}}) d\bar{\boldsymbol{w}},
\end{aligned} \tag{3.9}$$

where $g(\bar{\boldsymbol{w}})$ is the p.d.f. of $\boldsymbol{w}$ given in (3.6). 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) by Assumption 3.5. □

Note that the proof of Theorem 3.6 illustrates that the success probability of $(\mathrm{RP}\mathcal{X})$, though positive, depends on the choice of $\boldsymbol{p}$[7]. Next, under additional problem assumptions, we derive lower bounds on the success probability of $(\mathrm{RP}\mathcal{X})$ that are independent of $\boldsymbol{p}$ and/or quantifiable.

3.2 Quantifying the success probability of $(\mathrm{RP}\mathcal{X})$ 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 3.7. *Let Assumption 2.2 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. Assume that $\boldsymbol{p}_\top \neq \boldsymbol{x}^*_\top$, where the subscript represents the Euclidean projection on the effective subspace. Then*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] \geq \mathbb{P}[-\boldsymbol{1} - \boldsymbol{p}_{d_e+1:D} \leq \boldsymbol{w} \leq \boldsymbol{1} - \boldsymbol{p}_{d_e+1:D}], \tag{3.10}$$

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

[7] When $\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\| \to 0$, the multivariate t-distribution in Corollary 3.4 becomes degenerate. Thus it is challenging to derive a lower bound on the integral (3.9) that is uniformly bounded away from zero with respect to $\boldsymbol{p}$.

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

$$\boldsymbol{x}_\top^* = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}^* = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \boldsymbol{x}^* = \begin{pmatrix} \boldsymbol{x}_{1:d_e}^* \\ \boldsymbol{0} \end{pmatrix}. \tag{3.11}$$

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 3.4 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 (3.10). □

Note that the right-hand side of (3.10) can be written as the integral of the p.d.f. of $\boldsymbol{w}$ over the hyperrectangular region $-\boldsymbol{1} - \boldsymbol{p}_{d_e+1:D} \leq \boldsymbol{w} \leq \boldsymbol{1} - \boldsymbol{p}_{d_e+1:D}$. 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 Appendix D.

Theorem 3.8. *Let Assumption 2.2 hold with $\boldsymbol{U} = [\boldsymbol{I}_{d_e}\ \boldsymbol{0}]^T$ and $\boldsymbol{V} = [\boldsymbol{0}\ \boldsymbol{I}_{D-d_e}]^T$. Let d_e and d be fixed, and let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix. For all $\boldsymbol{p} \in \mathcal{X}$, we have*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] \geq \tau > 0, \tag{3.12}$$

where τ satisfies

$$\tau = \Theta\left(\frac{\log(D - d_e + 1)^{\frac{d-1}{2}}}{2^{D-d_e} \cdot (D - d_e + 1)^{d_e}}\right) \text{ as } D \to \infty, \tag{3.13}$$

and the constants in $\Theta(\cdot)$ depend only on d_e and d.

Proof. See Appendix D. □

The next result shows that, in the particular case when $\boldsymbol{p} = \boldsymbol{0}$, the center of the full-dimensional domain $\mathcal{X}$, the probability of success decreases at worst algebraically[8] with the ambient dimension D.

Theorem 3.9. *Let Assumption 2.2 hold with $\boldsymbol{U} = [\boldsymbol{I}_{d_e}\ \boldsymbol{0}]^T$ and $\boldsymbol{V} = [\boldsymbol{0}\ \boldsymbol{I}_{D-d_e}]^T$. Let d_e and d be fixed, and let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix. Let $\boldsymbol{p} = \boldsymbol{0}$. Then*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] \geq \tau_{\boldsymbol{0}} > 0, \tag{3.14}$$

where

$$\tau_{\boldsymbol{0}} = \Theta\left(\frac{\log(D - d_e + 1)^{\frac{d-1}{2}}}{(D - d_e + 1)^{d_e}}\right) \text{ as } D \to \infty, \tag{3.15}$$

and where the constants in $\Theta(\cdot)$ depend only on d_e and d.

Proof. See Appendix D. □

[8]This simplification is due to the fact that when $\boldsymbol{p} = \boldsymbol{0}$, the factor 2^{D-d_e} in the denominator of (3.13) disappears.

Remark 3.10. Unlike Theorem 3.6, the above result does not require Assumption 3.5. In this specific case, as the effective subspace is aligned with the coordinate axes, Assumption 3.5 is satisfied. The latter follows from $\bar{G}^* = \{\boldsymbol{g} \in \mathbb{R}^{D-d_e} : -\mathbf{1} \leq \boldsymbol{x}_\top^* + \boldsymbol{V}\boldsymbol{g} \leq \mathbf{1}\} = \{\boldsymbol{g} \in [-1,1]^{D-d_e}\}$, 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 3.7.

Remark 3.11. 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 (3.10) holds with equality if $d = d_e$ and $G = G^*$. Our numerical experiments in Section 5 also clearly illustrate that the success probability decreases with growing problem dimension D.

Remark 3.12. 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 (D.2) in [56]. 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 (3.13) and (3.15) for such choices of d_e and d, refer to [50, 56].

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

As mentioned in the last paragraph of Section 3.1, 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 3.13. 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

$$G_\epsilon := \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) \leq f^* + \epsilon\} \tag{3.16}$$

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 3.14. *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 3.15. *Suppose that Assumption 3.5 and Assumption 3.14 hold, and let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix and $\epsilon > 0$, an accuracy tolerance. Then there exists a constant $\tau_\epsilon > 0$ such that, for all $\boldsymbol{p} \in \mathcal{X}$,*

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] \geq \tau_\epsilon. \tag{3.17}$$

Proof. Assumption 3.5 implies that there exists a global minimizer $\boldsymbol{x}^* \in \mathcal{X}$, with corresponding sets G^* (Definition 2.3) and $\bar{G}^*$ in (2.4) such that $\text{Vol}(\bar{G}^*) > 0$. 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 3.14, $|f(\boldsymbol{p}) - f^*| = |f(\boldsymbol{p}_\top) - f(\boldsymbol{x}_\top^*)| \leq L\|\boldsymbol{x}_\top^* - \boldsymbol{p}_\top\| \leq \epsilon$. Thus $\boldsymbol{p} \in G_\epsilon$ and, hence, $\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] = 1$.

Otherwise, $\boldsymbol{p} \in \mathcal{X} \setminus N_{\epsilon/L}(G^*)$. Using the proof of Theorem 3.6, we have

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] \geq \mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] \geq \int_{-\boldsymbol{V}^T\boldsymbol{p}+\bar{G}^*} g(\bar{\boldsymbol{w}})d\bar{\boldsymbol{w}}, \tag{3.18}$$

where $g(\bar{\boldsymbol{w}})$ is the p.d.f. of $\boldsymbol{w}$ given by (3.6), and where the first inequality is due to the fact that $(\mathrm{RP}\mathcal{X})$ being successful implies that $(\mathrm{RP}\mathcal{X})$ is ϵ-successful (by letting $\epsilon := 0$ in Definition 3.13). To prove (3.17), 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

$$\frac{\epsilon}{L} < \|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|_2 = \|\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{x}^* - \boldsymbol{p})\|_2 \leq \|\boldsymbol{U}\boldsymbol{U}^T\|_2 \cdot \|\boldsymbol{x}^* - \boldsymbol{p}\|_2 \leq 2\sqrt{D}, \tag{3.19}$$

where the last inequality follows from $\|\boldsymbol{U}\boldsymbol{U}^T\|_2 = 1$, since $\boldsymbol{U}$ has orthonormal columns, and from $-\mathbf{2} \leq \boldsymbol{x}^* - \boldsymbol{p} \leq \mathbf{2}$ 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

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

and, hence,

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

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,

$$\|\bar{\boldsymbol{w}}\|_2 = \|\boldsymbol{V}\bar{\boldsymbol{w}}\|_2 \leq \sqrt{D}\|\boldsymbol{V}\bar{\boldsymbol{w}}\|_\infty \leq \sqrt{D}(1 + 2\sqrt{D}) \leq 3D. \tag{3.20}$$

By combining (3.6), (3.19) and (3.20), we finally obtain

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

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

4 The X-REGO algorithm and its global convergence

In the case of random embeddings for unconstrained global optimization [9], the success probability of the reduced problem is independent of the ambient dimension [9]. However, in the constrained case of problem (P), the analysis in Section 3 shows that the probability of success of the reduced problem $(\mathrm{RP}\mathcal{X})$ decreases with D. 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 1). 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.

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).

Algorithm 1 $\mathcal{X}$-Random Embeddings for Global Optimization (X-REGO) applied to (P)

1: Initialize d and $\boldsymbol{p}^0 \in \mathcal{X}$
2: **for** $k \geq 1$ until termination **do**
3: Draw $\tilde{\boldsymbol{A}}^k$, a realization of the $D \times d$ Gaussian matrix $\boldsymbol{A}$
4: Calculate $\tilde{\boldsymbol{y}}^k$ by solving approximately and possibly, probabilistically,

$$\begin{aligned} \tilde{f}^k_{min} = \min_{\boldsymbol{y} \in \mathbb{R}^d} \; & f(\tilde{\boldsymbol{A}}^k \boldsymbol{y} + \tilde{\boldsymbol{p}}^{k-1}) \\ \text{subject to } & \tilde{\boldsymbol{A}}^k \boldsymbol{y} + \tilde{\boldsymbol{p}}^{k-1} \in \mathcal{X} \end{aligned} \qquad (\widetilde{\text{RP}\mathcal{X}^k})$$

5: Let

$$\tilde{\boldsymbol{x}}^k := \tilde{\boldsymbol{A}}^k \tilde{\boldsymbol{y}}^k + \tilde{\boldsymbol{p}}^{k-1} \qquad (4.1)$$

6: Choose (deterministically or randomly) $\tilde{\boldsymbol{p}}^k \in \mathcal{X}$
7: **end for**

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 $\tilde{f}^k_{min} = f^k_{min}(\boldsymbol{\omega}^k)$, for $k \geq 1$, that are realizations of the random variables $\boldsymbol{x}^k$, $\boldsymbol{y}^k$ and f^k_{min}, respectively. Each iteration of X-REGO solves – approximately and possibly, with a certain probability – a realization $(\widetilde{\text{RP}\mathcal{X}^k})$ of the random problem

$$\begin{aligned} f^k_{min} = \min_{\boldsymbol{y}} \; & f(\boldsymbol{A}^k \boldsymbol{y} + \boldsymbol{p}^{k-1}) \\ \text{subject to } & \boldsymbol{A}^k \boldsymbol{y} + \boldsymbol{p}^{k-1} \in \mathcal{X}. \end{aligned} \qquad (\text{RP}\mathcal{X}^k)$$

To calculate $\tilde{\boldsymbol{y}}^k$, $(\widetilde{\text{RP}\mathcal{X}^k})$ 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 $(\widetilde{\text{RP}\mathcal{X}^k})$ (at least) with a certain probability.

Several variants of X-REGO can be obtained by specific choices of the random variable $\boldsymbol{p}^k$ (assumed throughout the paper to have support contained in $\mathcal{X}$). A first possibility consists in simply defining $\boldsymbol{p}^k$ as a random variable with support $\{\boldsymbol{0}\}$, so that $\tilde{\boldsymbol{p}}^k = \boldsymbol{0}$ for all k. It is also possible to preserve the progress achieved so far by defining $\boldsymbol{p}^k = \boldsymbol{x}^k_{opt}$, where

$$\boldsymbol{x}^k_{opt} := \arg\min\{f(\boldsymbol{x}^1), f(\boldsymbol{x}^2), \ldots, f(\boldsymbol{x}^k)\}, \qquad (4.2)$$

the random variable corresponding to the best point found over the k first embeddings. We compare numerically several choices of $\boldsymbol{p}$ on benchmark functions in Section 5.

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 for now.

4.1 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 (3.16). We show that $\boldsymbol{x}^k_{opt}$ in (4.2) converges to G_ϵ almost surely as $k \to \infty$ (see Theorem 4.7).

Intuitively, our proof relies on the fact that any vector $\tilde{\boldsymbol{x}}^k$ defined in (4.1) 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 3.13[9], namely,

$$f^k_{min} \leq f^* + \epsilon - \lambda; \tag{4.3}$$

(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,

$$f(\boldsymbol{A}^k\boldsymbol{y}^k + \boldsymbol{p}^{k-1}) \leq f^k_{min} + \lambda \tag{4.4}$$

holds (at least) with a certain probability. We introduce two additional random variables that capture the conditions in (a) and (b) above,

$$R^k = \mathbb{1}\{(\text{RP}\mathcal{X}^k) \text{ is } (\epsilon - \lambda)\text{-successful in the sense of (4.3)}\}, \tag{4.5}$$

$$S^k = \mathbb{1}\{(\text{RP}\mathcal{X}^k) \text{ is solved to accuracy } \lambda \text{ in the sense of (4.4)}\}, \tag{4.6}$$

where $\mathbb{1}$ is the usual indicator function for an event.

Let $\mathcal{F}^k = \sigma(\boldsymbol{A}^1, \ldots, \boldsymbol{A}^k, \boldsymbol{y}^1, \ldots, \boldsymbol{y}^k, \boldsymbol{p}^0, \ldots, \boldsymbol{p}^k)$ be the σ-algebra generated by the random variables $\boldsymbol{A}^1, \ldots, \boldsymbol{A}^k, \boldsymbol{y}^1, \ldots, \boldsymbol{y}^k, \boldsymbol{p}^0, \ldots, \boldsymbol{p}^k$ (a mathematical concept that represents the history of the X-REGO algorithm as well as its randomness until the kth embedding)[10], with $\mathcal{F}^0 = \sigma(\boldsymbol{p}^0)$. We also construct an 'intermediate' σ-algebra, namely,

$$\mathcal{F}^{k-1/2} = \sigma(\boldsymbol{A}^1, \ldots, \boldsymbol{A}^{k-1}, \boldsymbol{A}^k, \boldsymbol{y}^1, \ldots, \boldsymbol{y}^{k-1}, \boldsymbol{p}^0, \ldots, \boldsymbol{p}^{k-1}),$$

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[11], and R^k is also $\mathcal{F}^{k-1/2}$-measurable; thus they are well-defined random variables.

Remark 4.1. The random variables $\boldsymbol{A}^1, \ldots, \boldsymbol{A}^k$, $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^k$, $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k$, $\boldsymbol{p}^0, \boldsymbol{p}^1, \ldots, \boldsymbol{p}^k$, $R^1, \ldots, R^k$, $S^1, \ldots, S^k$ are $\mathcal{F}^k$-measurable since $\mathcal{F}^0 \subseteq \mathcal{F}^1 \subseteq \cdots \subseteq \mathcal{F}^k$. Also, $\boldsymbol{A}^1, \ldots, \boldsymbol{A}^k$, $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^{k-1}$, $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{k-1}$, $\boldsymbol{p}^0, \boldsymbol{p}^1, \ldots, \boldsymbol{p}^{k-1}$, $R^1, \ldots, R^k$, $S^1, \ldots, S^{k-1}$ 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.

---

[9] 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 ($\widetilde{\text{RP}\mathcal{X}^k}$).

[10] A similar setup for random iterates of probabilistic models can be found in [2, 10].

[11] 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 $\mathcal{F}^{k-1/2}$.

Assumption 4.2. *There exists $\rho \in (0,1]$ such that, for all $k \geq 1$,*[12]

$$\mathbb{P}[S^k = 1 | \mathcal{F}^{k-1/2}] = \mathbb{E}[S^k | \mathcal{F}^{k-1/2}] \geq \rho,$$

i.e., with (conditional) probability at least $\rho > 0$, the solution $\boldsymbol{y}^k$ of (RP$\mathcal{X}^k$) *satisfies* (4.4).

Remark 4.3. If a deterministic (global optimization) algorithm is used to solve ($\widetilde{\text{RP}\mathcal{X}^k}$), then S^k is always $\mathcal{F}_k^{k-1/2}$-measurable and Assumption 4.2 is equivalent to $S^k \geq \rho$. Since S^k is an indicator function, this further implies that $S^k \equiv 1$, provided a sufficiently large computational budget is available.

The results of Section 3 provide a lower bound on the (conditional) probability of the reduced problem (RP$\mathcal{X}^k$) to be $(\epsilon - \lambda)$-successful, with the consequence given in the first part of the next Corollary.

Corollary 4.4. *If Assumptions 3.5 and 3.14 hold, then*

$$\mathbb{E}[R^k | \mathcal{F}^{k-1}] \geq \tau, \quad \text{for} \quad k \geq 1. \tag{4.7}$$

If Assumption 4.2 holds, then

$$\mathbb{E}[R^k S^k | \mathcal{F}^{k-1/2}] \geq \rho R^k, \quad \text{for} \quad k \geq 1. \tag{4.8}$$

Proof. Recall that the support of the random variable $\boldsymbol{p}^k$ is contained in $\mathcal{X}$. For each embedding, we apply Theorem 3.15 (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 $k \geq 1$. Then, in terms of conditional expectation, we have $\mathbb{E}[R^k | \mathcal{F}^{k-1}] = 1 \cdot \mathbb{P}[R^k = 1 | \mathcal{F}^{k-1}] + 0 \cdot \mathbb{P}[R^k = 0 | \mathcal{F}^{k-1}] \geq \tau$.

If Assumption 4.2 holds, then $\mathbb{E}[R^k S^k | \mathcal{F}^{k-1/2}] = R^k \mathbb{E}[S^k | \mathcal{F}^{k-1/2}] \geq \rho R^k$, where the equality follows from the fact that R^k is $\mathcal{F}^{k-1/2}$-measurable (see [18, Theorem 4.1.14]). □

4.1.1 Global convergence proof

A useful property is given next.

Lemma 4.5. *Let Assumptions 3.5, 3.14 and 4.2 hold. Then, for $K \geq 1$, we have*

$$\mathbb{P}\Big[\bigcup_{k=1}^{K} \Big\{ \{R^k = 1\} \cap \{S^k = 1\} \Big\} \Big] \geq 1 - (1 - \tau\rho)^K.$$

Proof. We define an auxiliary random variable, $J^K := \mathbb{1}\left(\bigcup_{k=1}^{K} \big\{ \{R^k = 1\} \cap \{S^k = 1\} \big\} \right)$. Note that $J^K = 1 - \prod_{k=1}^{K} (1 - R^k S^k)$. We have

$$\begin{aligned}
\mathbb{P}\Big[\bigcup_{k=1}^{K} \Big\{ \{R^k = 1\} \cap \{S^k = 1\} \Big\} \Big] = \mathbb{E}[J^K] &= 1 - \mathbb{E}\Big[\prod_{k=1}^{K} (1 - R^k S^k) \Big] \\
&\overset{(*)}{=} 1 - \mathbb{E}\Big[\mathbb{E}\Big[\prod_{k=1}^{K} (1 - R^k S^k) \Big| \mathcal{F}^{K-1/2} \Big] \Big] \\
&\overset{(\circ)}{=} 1 - \mathbb{E}\Big[\prod_{k=1}^{K-1} (1 - R^k S^k) \cdot \mathbb{E}\big[1 - R^K S^K | \mathcal{F}^{K-1/2} \big] \Big] \\
&\geq 1 - \mathbb{E}\Big[(1 - \rho R^K) \cdot \prod_{k=1}^{K-1} (1 - R^k S^k) \Big]
\end{aligned}$$

[12] The equality in the displayed equation follows from $\mathbb{E}[S^k | \mathcal{F}^{k-1}] = 1 \cdot \mathbb{P}[S^k = 1 | \mathcal{F}^{k-1}] + 0 \cdot \mathbb{P}[S^k = 0 | \mathcal{F}^{k-1}]$.

$$\overset{(*)}{=} 1 - \mathbb{E}\Big[\mathbb{E}\Big[(1-\rho R^K)\cdot\prod_{k=1}^{K-1}(1-R^kS^k)\Big|\mathcal{F}^{K-1}\Big]\Big]$$
$$\overset{(\circ)}{=} 1 - \mathbb{E}\Big[\prod_{k=1}^{K-1}(1-R^kS^k)\cdot\mathbb{E}\big[1-\rho R^K|\mathcal{F}^{K-1}\big]\Big]$$
$$\geq 1 - (1-\tau\rho)\cdot\mathbb{E}\Big[\prod_{k=1}^{K-1}(1-R^kS^k)\Big],$$

where

- $(*)$ follow from the tower property of conditional expectation (see (4.1.5) in [18]),
- $(\circ)$ 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 [18]),
- the inequalities follow from (4.8) and (4.7), 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 4.1.

Lemma 4.6. *Suppose Assumptions 3.5, 3.14 and 4.2 hold. Then*

$$\{R^k=1\}\cap\{S^k=1\}\subseteq\{\boldsymbol{x}^k\in G_\epsilon\}.$$

Proof. By Definition 3.13, 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

$$f(\boldsymbol{A}^k\boldsymbol{y}^k_{int}+\boldsymbol{p}^{k-1})\leq f^*+\epsilon-\lambda. \tag{4.9}$$

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

$$f^k_{min}\leq f(\boldsymbol{A}^k\boldsymbol{y}^k_{int}+\boldsymbol{p}^{k-1}). \tag{4.10}$$

Then, for $\boldsymbol{x}^k$, (4.4) 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 (4.10) and (4.9), respectively. This shows that $\boldsymbol{x}^k\in G_\epsilon$. □

Theorem 4.7 (Global convergence)**.** *Suppose Assumptions 3.5, 3.14 and 4.2 hold. Then*

$$\lim_{k\to\infty}\mathbb{P}[\boldsymbol{x}^k_{opt}\in G_\epsilon]=\lim_{k\to\infty}\mathbb{P}[f(\boldsymbol{x}^k_{opt})\leq f^*+\epsilon]=1$$

where $\boldsymbol{x}^k_{opt}$ and G_ϵ are defined in (4.2) *and* (3.16)*, respectively.*

Furthermore, for any $\xi\in(0,1)$,

$$\mathbb{P}[\boldsymbol{x}^k_{opt}\in G_\epsilon]=\mathbb{P}[f(\boldsymbol{x}^k_{opt})\leq f^*+\epsilon]\geq\xi \text{ for all } k\geq K_\xi, \tag{4.11}$$

where $K_\xi:=\left\lceil\dfrac{|\log(1-\xi)|}{\tau\rho}\right\rceil$.

Proof. Lemma 4.6 and the definition of $\boldsymbol{x}^k_{opt}$ in (4.2) provide

$$\{R^k = 1\} \cap \{S^k = 1\} \subseteq \{\boldsymbol{x}^k \in G_\epsilon\} \subseteq \{\boldsymbol{x}^k_{opt} \in G_\epsilon\}$$

for $k = 1, 2, \ldots, K$ and for any integer $K \geq 1$. Hence,

$$\bigcup_{k=1}^{K} \{R^k = 1\} \cap \{S^k = 1\} \subseteq \bigcup_{k=1}^{K} \{\boldsymbol{x}^k_{opt} \in G_\epsilon\}. \tag{4.12}$$

Note that the sequence $\{f(\boldsymbol{x}^1_{opt}), f(\boldsymbol{x}^2_{opt}), \ldots, f(\boldsymbol{x}^K_{opt})\}$ is monotonically decreasing. Therefore, if $\boldsymbol{x}^k_{opt} \in G_\epsilon$ for some $k \leq K$ then $\boldsymbol{x}^i_{opt} \in G_\epsilon$ for all $i = k, \ldots, K$; and so the sequence $(\{\boldsymbol{x}^k_{opt} \in G_\epsilon\})_{k=1}^K$ is an increasing sequence of events. Hence,

$$\bigcup_{k=1}^{K} \{\boldsymbol{x}^k_{opt} \in G_\epsilon\} = \{\boldsymbol{x}^K_{opt} \in G_\epsilon\}. \tag{4.13}$$

From (4.13) and (4.12), we have for all $K \geq 1$,

$$\mathbb{P}[\{\boldsymbol{x}^K_{opt} \in G_\epsilon\}] \geq \mathbb{P}\Big[\bigcup_{k=1}^{K} \{R^k = 1\} \cap \{S^k = 1\}\Big] \geq 1 - (1 - \tau\rho)^K, \tag{4.14}$$

where the second inequality follows from Lemma 4.5. Finally, passing to the limit with K in (4.14), we deduce $1 \geq \lim_{K\to\infty} \mathbb{P}[\{\boldsymbol{x}^K_{opt} \in G_\epsilon\}] \geq \lim_{K\to\infty} \left[1 - (1 - \tau\rho)^K\right] = 1$, as required.

Note that if

$$1 - (1 - \tau\rho)^k \geq \xi \tag{4.15}$$

then (4.14) implies $\mathbb{P}[\boldsymbol{x}^k_{opt} \in G_\epsilon] \geq \xi$. Since (4.15) is equivalent to $k \geq \dfrac{\log(1-\xi)}{\log(1-\tau\rho)}$, (4.15) holds for all $k \geq K_\xi$ since $K_\xi \geq \dfrac{\log(1-\xi)}{\log(1-\tau\rho)}$. □

Remark 4.8. Crucially, we note that X-REGO (Algorithm 1) is a generic framework that can be applied to a general, continuous objective f in (P). Furthermore, the convergence result in Theorem 4.7 also continues to hold in this general case provided (4.7) can be shown to hold; this is where we crucially use the special structure of low effective dimensionality of the objective that we investigate in this paper.

Remark 4.9. 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 (4.4). 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.

Quantifiable rates of convergence when the effective subspace is aligned with coordinate axes Using the estimates for τ in Theorem 3.8, 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 4.10. *Suppose Assumption 2.2 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 4.2. Let* $\xi \in (0,1)$*, and* d_e *and* d *be fixed. Then* (4.11) *holds with*

$$K_\xi = \frac{|\log(1-\xi)|}{\rho} O\left(\frac{2^{D-d_e} \cdot (D - d_e + 1)^{d_e}}{\log(D - d_e + 1)^{\frac{d-1}{2}}}\right) \text{ as } D \to \infty. \tag{4.16}$$

If $\boldsymbol{p}^k = \mathbf{0}$ *for* $k \geq 0$*, then* (4.11) *holds with*

$$K_\xi = \frac{|\log(1-\xi)|}{\rho} O\left(\frac{(D-d_e+1)^{d_e}}{\log(D-d_e+1)^{\frac{d-1}{2}}}\right) \text{ as } D \to \infty. \tag{4.17}$$

Proof. Firstly, note our remark regarding assumptions below. The result follows from Theorem 4.7, (3.13) and (3.15). □

Remark 4.11. Assumptions 3.5 and 3.14 were required to prove Theorem 3.15 and, consequently, (4.7). If the effective subspace is aligned with coordinate axes, we no longer need Assumptions 3.14 and 3.5 to prove (4.7). In this case, (4.7) follows from Theorem 3.8, together with the fact that (RP$\mathcal{X}^k$) being successful implies (RP$\mathcal{X}^k$) is ϵ-successful for any $\epsilon \geq 0$.

5 Numerical experiments

5.1 Setup

Algorithms. We test different variants of Algorithm 1 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 N-REGO). Then, if $|f(\boldsymbol{A}^k\boldsymbol{y}^k + \boldsymbol{p}^{k-1}) - f(\boldsymbol{p}^{k-1})| > \gamma$ 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: $\boldsymbol{p}^k := \mathbf{0}$ for all k.

- Local Nonadaptive X-REGO (LN-REGO). In X-REGO, the low-dimensional problem ($\widetilde{\text{RP}\mathcal{X}^k}$) 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 ($\widetilde{\text{RP}\mathcal{X}^k}$) (or the original problem (P) in the no-embedding case), namely, DIRECT ([22, 25, 33]), BARON ([46, 49]) and KNITRO ([8]).

DIRECT([25, 33, 22]) version 4.0 (DIviding RECTangles) is a deterministic[13] global optimization solver, that does not require information about the gradient nor about the Lipschitz constant.

BARON([46, 49]) version 17.10.10 (Branch-And-Reduce Optimization Navigator) is a state-of-the-art branch- and-bound type global optimization solver for nonlinear and mixed-integer programs, that is highly competitive [42]. However, it accepts only a few (general) classes of functions (e.g., no trigonometric functions, no black box functions).

KNITRO([8]) version 10.3.0 is a large-scale nonlinear local optimization solver that makes use of objective derivatives. KNITRO has a multi-start feature, referred here as mKNITRO, allowing it to aim for global minimizers.

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

We refer to [9] for a detailed description of the solvers. We test A-REGO and N-REGO using DIRECT, BARON and mKNITRO and test LA-REGO and LN-REGO using only local KNITRO, with no multi-start.

Test set. The methodology of these constructions is given in [9, 54] and summarized here in Appendix E. Our synthetic test set contains 19 D-dimensional functions with low effective dimension, with $D = 10, 100$ and 1000. We construct these high-dimensional functions from 19 global optimization problems (Table 3, of dimensions 2–6) with known global minima [20, 27, 5], some of which are in the Dixon-Szego test set [14]. The construction process consists in artificially adding coordinates to the original functions, and then applying a rotation to ensure that the effective subspace is not aligned with the coordinate axes.

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[14]

$$f(\widetilde{\boldsymbol{A}}^k \tilde{\boldsymbol{y}}^k + \tilde{\boldsymbol{p}}^{k-1}) - f^* \leq \epsilon = 10^{-3}. \tag{5.1}$$

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 1.

Remark 5.1. 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.

We compare the results using performance profiles (Dolan and Moré, [16]), which measure the proportion of problems solved by the algorithm in less than a given budget defined based on the best performance among the algorithms considered. More precisely, for each solver (BARON, DIRECT and KNITRO), and for each algorithm $\mathcal{A}$ (the above-mentioned variants of X-REGO and 'no-embedding'), we record $\mathcal{N}_p(\mathcal{A})$, the computational cost (see Table 1) of running algorithm $\mathcal{A}$ to solve problem p within accuracy ϵ. Let $\mathcal{N}_p^*$ be the minimum computational cost required for problem p by any algorithm $\mathcal{A}$. The performance (probability) of algorithm $\mathcal{A}$ on the problem set $\mathcal{P}$ is defined as

$$\pi_{\mathcal{A}}(\alpha) = \frac{|\{p \in \mathcal{P} : \mathcal{N}_p(\mathcal{A}) \leq \alpha \mathcal{N}_p^*\}|}{|\mathcal{P}|},$$

with performance ratio $\alpha \geq 1$. As each experiment involving random embeddings is repeated five times, we obtain five curves for the corresponding algorithm-solver pairs.

5.2 Numerical results

DIRECT: Figure 2 compares the adaptive and non-adaptive random embedding algorithms (A-REGO and N-REGO) to the no-embedding framework, when using the DIRECT solver

[14]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 1: The table outlines the experimental setup for the solvers, used both in the 'no-embedding' algorithm and for solving the low-dimensional problem ($\widetilde{\text{RP}\mathcal{X}^k}$) (as usually, f denotes the D-dimensional function to minimize, f^* is its global minimum, and ϵ in (5.1) is set to 10^{-3}). At each internal iteration, DIRECT stores f_D^* — the minimal value of f found so far, while BARON stores f_B^U and f_B^L — the smallest upper bound and largest lower bound found found so far. Note that, for BARON, $f_B^U = f(\boldsymbol{x}^k)$ in ($\widetilde{\text{RP}\mathcal{X}^k}$).

	DIRECT	BARON	mKNITRO	KNITRO
Measure of computational cost	function evaluations	CPU seconds	function evaluations	function evaluations
Max. budget to solve ($\widetilde{\text{RP}\mathcal{X}^k}$)	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 ($\widetilde{\text{RP}\mathcal{X}^k}$)	Terminate either on budget or if $f_D^* \leq f^* + \epsilon$	Terminate either on budget or if f_B^U and f_B^L satisfy $f_B^U \leq f_B^L + \epsilon$	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_B^U satisfies $f_B^U \leq f^* + \epsilon$	Same as above	Not applicable
Additional options for ($\widetilde{\text{RP}\mathcal{X}^k}$)	`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

for the reduced problem ($\widetilde{\text{RP}\mathcal{X}^k}$) (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 2.

BARON: Figure 3 compares A-REGO and N-REGO to the no-embedding framework, when using BARON to solve the reduced problem ($\widetilde{\text{RP}\mathcal{X}^k}$). 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-adaptive one. Table 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 ($\widetilde{\text{RP}\mathcal{X}^k}$), is given in Figure 4. 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 4 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 2, which indicates that

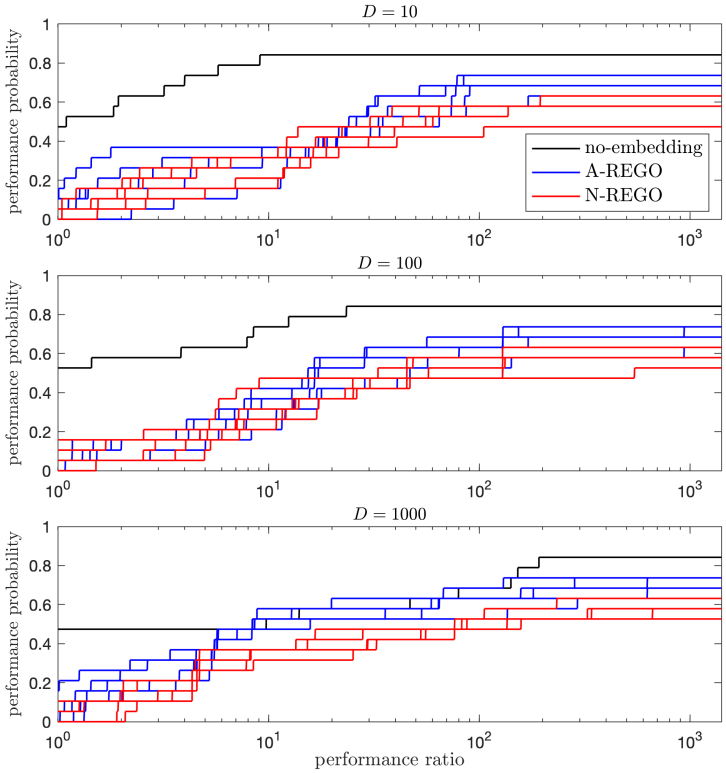

Figure 2: Comparison between X-REGO variants and 'no-embedding', using DIRECT to solve the subproblem ($\widetilde{\text{RP}\mathcal{X}^k}$).

Table 2: Median number of function evaluations or CPU time spent by each algorithm-solver pair.

	DIRECT (fun. evals)			BARON (CPU time)			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
N-REGO	63093	300484	300532	0.82	3.00	21.51	1582	3606	8766
LA-REGO	–	–	–	–	–	–	368	631	2564
LN-REGO	–	–	–	–	–	–	220	763	704

the median number of function evaluations increases significantly for LA-REGO while for LN-REGO, it actually decreases.

Conclusions to numerical experiments The numerical experiments presented in this paper indicate that, as expected, the X-REGO algorithm is mostly beneficial for high-dimensional problems, when D is large. In this setting, X-REGO variants paired with the BARON and mKNITRO solvers outperform the 'no-embedding' approach, of applying these solvers directly to the problems. It is less obvious to decide which variant of X-REGO is best, but it seems that, at least on the problem set considered, the local variants outperform the global ones.

6 Conclusions and future work

We studied a generic global optimization framework, X-REGO, that relies on multiple random embeddings, for bound-constrained global optimization of functions with low effective dimension-

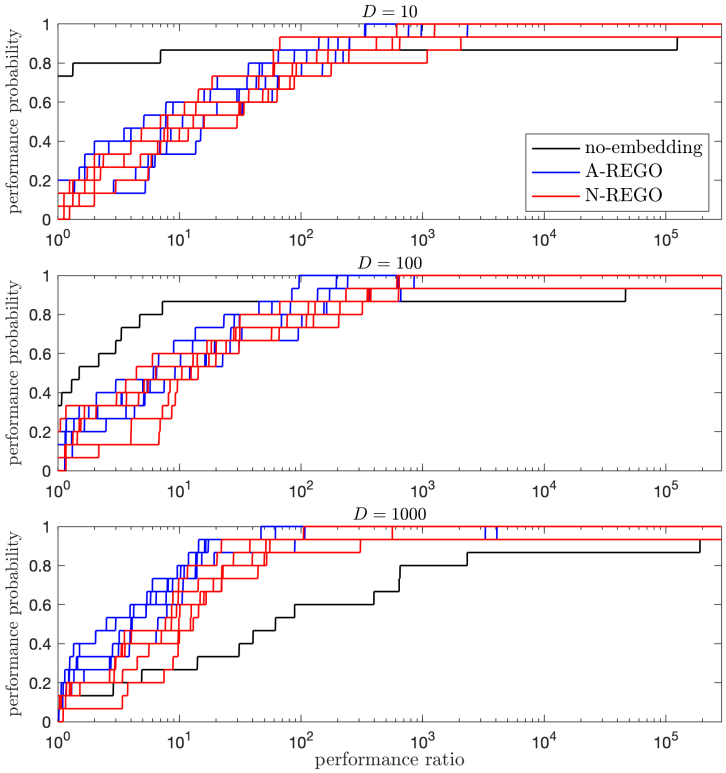

Figure 3: Comparison between X-REGO variants and 'no-embedding', using BARON to solve the subproblem ($\widetilde{\mathrm{RP}\mathcal{X}^k}$).

ality. For each random subspace, a lower-dimensional bound-constrained subproblem is solved, using a global or even local algorithm. Theoretical guarantees of convergence and encouraging numerical experiments are presented, which are particularly quantified in terms of their problem dimension dependence for the case when the effective subspace is aligned with the coordinate axes. We note that the X-REGO algorithmic framework (Algorithm 1) can be applied to a general, continuous objective f in (P) as the effective dimensionality assumption is not used; furthermore, our main global convergence result continues to hold under some assumptions (see Remark 4.8).

Our analysis relies on the assumption that the dimension of the random subspace is larger than the effective dimension. As the latter may be unknown in practice, this is a strong prerequisite. One possibility is to estimate the effective dimension d_e numerically, as in [47]. Otherwise, one may consider extending the theoretical analysis in this paper to the case $d \leq d_e$. A relevant recent reference is [34], where Kirschner et al. proved global convergence of an algorithm similar to A-REGO, but using one-dimensional subspaces, within the framework of Bayesian optimization.

References

[1] T. Amdeberhan and V. H. Moll, editors. *Tapas in Experimental Mathematics*, Contemporary Mathematics 457, 2008. American Mathematical Society.

[2] 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.

[3] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(1):281–305, 2012.

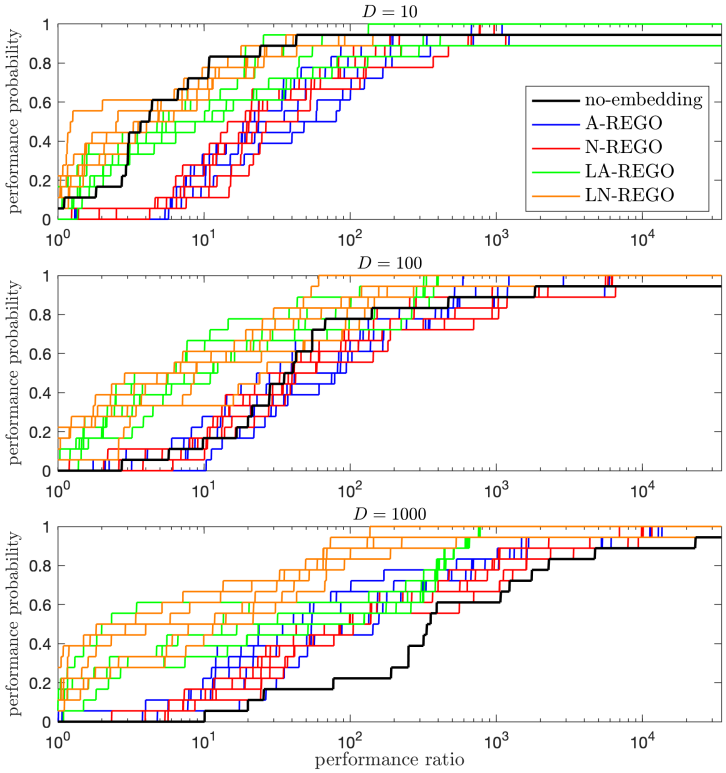

Figure 4: Comparison between X-REGO variants and 'no-embedding', using KNITRO to solve the subproblem ($\widetilde{\text{RP}\mathcal{X}^k}$).

[4] J. M. Bernardo and A. F.M. Smith. *Bayesian theory*. Wiley, 2000.

[5] D. Bingham. Virtual library of simulation experiments: test functions and datasets. https://www.sfu.ca/~ssurjano/, 2013. Accessed: 2017-01-27.

[6] 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.

[7] 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.

[8] R. H. Byrd, J. Nocedal, and R. A. Waltz. *Knitro: An Integrated Package for Nonlinear Optimization*, pages 35–59. Springer US, Boston, MA, 2006.

[9] 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.

[10] C. Cartis and K. Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Math. Program.*, 169(2):337–375, 2018.

[11] 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.

[12] P. Constantine. *Active Subspaces*. SIAM, Philadelphia, PA, 2015.

[13] 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.

[14] L.C.W. Dixon and G.P. Szegö. *Towards Global Optimization*. Elseiver, New York, 1975.

[15] 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.

[16] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, 2002.

[17] C. W. Dunnett and M. Sobel. Approximations to the probability integral and certain percentage points of a multivariate analogue of student's t-distribution. *Biometrika*, 42(1/2):258–260, 1955.

[18] R. Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition, 2019.

[19] 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.

[20] P.A. Ernesto and U.P. Diliman. MVF—multivariate test functions library in C for unconstrained global optimization, 2005.

[21] K. Fang, S. Kotz, and K. W. Ng. *Symmetric multivariate and related distributions*. London: Chapman and Hall, 1990.

[22] D. E. Finkel. *Direct optimization algorithm user guide*, 2003. Available at http://www2.peq.coppe.ufrj.br/Pessoal/Professores/Arge/COQ897/Naturais/DirectUserGuide.pdf.

[23] 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.

[24] 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.

[25] J.M. Gablonsky and C.T. Kelley. A locally-biased form of the direct algorithm. *Journal of Global Optimization*, 21(1):27–37, 2001.

[26] 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.

[27] A. Gavana. Global optimization benchmarks and AMPGO. Available at http://infinity77.net/global_optimization/.

[28] R. Gower, D. Koralev, F. Lieder, and P. Richtárik. Rsn: Randomized subspace newton. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, NIPS'19, pages 616–625, 2019.

[29] A.K. Gupta and D.K. Nagar. *Matrix Variate Distributions*. New York: Chapman and Hall/CRC, 2000.

[30] A.K. Gupta and D. Song. Lp-norm spherical distribution. *Journal of Statistical Planning and Inference*, 60 (2):241–260, 1997.

[31] F. Hanzely, N. Doikov, P. Richtárik, and Y. Nesterov. Stochastic subspace cubic newton method. *arXiv preprint arXiv:2002.09526*, 2020.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] D. Kozak, S. Becker, A. Doostan, and L. Tenorio. Stochastic subspace descent. *arXiv e-prints*, page arXiv:1904.01145, 2019.

[37] P. M. Lee. *Bayesian Statistics: An Introduction.* John Wiley & Sons, 4th edition, 2012.

[38] B. Letham, R. Calandra, A. Rai, and E. Bakshy. Re-examining linear embeddings for high-dimensional bayesian optimization, 2020. URL https://openreview.net/forum?id=SJgn3lBtwH.

[39] 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.

[40] 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.

[41] Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

[42] A. Neumaier, O. Shcherbina, W. Huyer, and T. Vinkó. A comparison of complete global optimization solvers. *Mathematical Programming*, 103(2):335–356, 2005.

[43] 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.

[44] H. Qian and Y. Yu. Solving high-dimensional multi-objective optimization problems with low effective dimensions. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'20, pages 875–881, 2020.

[45] 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.

[46] N. V. Sahinidis. *BARON 14.3.1: Global Optimization of Mixed-Integer Nonlinear Programs,* User's Manual, 2014.

[47] 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.

[48] S. Surjanovic and D. Bingham. Virtual library of simulation experiments: Test functions and datasets, 2013. Available at https://www.sfu.ca/~ssurjano/.

[49] M. Tawarmalani and N. V. Sahinidis. A polyhedral branch-and-cut approach to global optimization. *Mathematical Programming*, 103:225–249, 2005.

[50] N. Temme. *Asymptotic Methods for Integrals.* World Scientific, Singapore, 2014.

[51] H. Tran-The, S. Gupta, S. Rana, and S. Venkatesh. Trading convergence rate with computational budget in high dimensional bayesian optimization. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'20, 2020.

[52] 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.

[53] R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[54] 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.

[55] R. L. Wheeden. *Measure and integral : an introduction to real analysis.* Boca Raton: Chapman and Hall/CRC, 2nd edition, 2015.

[56] R. Wong. *Asymptotic Approximations of Integrals.* Society for Industrial and Applied Mathematics, 2001.

[57] 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.

A Technical definitions and results

A.1 Gaussian random matrices

Definition A.1. (Gaussian matrix, see [29, Definition 2.2.1]) A Gaussian (random) matrix is a matrix $\boldsymbol{A} = (a_{ij})$, where the entries $a_{ij} \sim \mathcal{N}(0,1)$ 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 that we use in the analysis; for a collection of results pertaining to Gaussian matrices and other related distributions refer to [29, 53].

Theorem A.2. *(see [29, Theorem 2.3.10]) Let $\boldsymbol{A}$ be a $D \times d$ Gaussian random matrix. If $\boldsymbol{U} \in \mathbb{R}^{D\times p}$, $D \geq p$, and $\boldsymbol{V} \in \mathbb{R}^{d\times q}$, $d \geq q$, are orthonormal, then $\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{V}$ is a Gaussian random matrix.*

Theorem A.3. *(see [29, Theorem 2.3.15]) Let $\boldsymbol{A}$ be a $D \times d$ Gaussian random matrix, and let $\boldsymbol{X} \in \mathbb{R}^{r\times D}$ and $\boldsymbol{Y} \in \mathbb{R}^{q\times D}$ be given matrices. Then, $\boldsymbol{X}\boldsymbol{A}$ and $\boldsymbol{Y}\boldsymbol{A}$ are independent if and only if $\boldsymbol{X}\boldsymbol{Y}^T = \boldsymbol{0}$.*

Theorem A.4. *(see [29, Theorem 3.2.1]) Let $\boldsymbol{A}$ be a $D\times d$ Gaussian random matrix with $D \geq d$. Then, the Wishart matrix $\boldsymbol{A}^T\boldsymbol{A}$ is positive definite, and hence nonsingular, with probability one.*

A.2 Other relevant probability distributions

Definition A.5 (Chi-squared distribution)**.** Given a collection $Z_1, Z_2, \ldots, Z_n$ of n independent standard normal variables, the random variable $W = Z_1^2 + Z_2^2 + \cdots Z_n^2$ is said to follow the chi-squared distribution with n degrees of freedom (see [37, A.2]). We denote this by $W \sim \chi_n^2$.

Theorem A.6. *(see [29, Theorem 3.3.12]) Let $\boldsymbol{M}$ be an $n \times l$ Gaussian matrix with $n \geq l$, $\boldsymbol{y}$ be an $l \times 1$ random vector distributed independently of $\boldsymbol{M}^T\boldsymbol{M}$, and $\mathbb{P}[\boldsymbol{y} \neq \boldsymbol{0}] = 1$. Then,*

$$\frac{\boldsymbol{y}^T\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{y}} \sim \chi_n^2$$

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

Definition A.7 (Inverse chi-squared distribution)**.** Given $X \sim \chi_n^2$, the random variable $Y = 1/X$ is said to follow the inverse chi-squared distribution with n degrees of freedom. We denote this by $Y \sim 1/\chi_n^2$ (see [37, A.5]).

Definition A.8 (Multivariate t-distribution)**.** An l-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 [29, Chapter 4])

$$f(\boldsymbol{t}) = \frac{1}{(\pi\nu)^{l/2}}\left[\frac{\Gamma\left(\frac{l+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right]\det(\boldsymbol{\Sigma})^{-1/2}\left(1 + \frac{1}{\nu}\boldsymbol{t}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{t}\right)^{-(l+\nu)/2}, \tag{A.1}$$

where Γ is the usual gamma function.

Definition A.9 (F-distribution)**.** Let $W_1 \sim \chi_{\nu_1}^2$ and $W_2 \sim \chi_{\nu_2}^2$ be independent. A random variable X is said to follow an F-distribution with degrees of freedom ν_1 and ν_2 if

$$X \sim \frac{W_1/\nu_1}{W_2/\nu_2}.$$

We denote this by $X \sim F(\nu_1, \nu_2)$. The p.d.f. of X is given by (see [37, A.19])

$$f(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{\nu_1/2-1}\left(1 + \frac{\nu_1}{\nu_2}x\right)^{-\frac{\nu_1+\nu_2}{2}} \text{ for } x > 0. \tag{A.2}$$

A.3 Additional relevant results

Lemma A.10. *[21, 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,*

$$\begin{pmatrix} f_1(\boldsymbol{x}) & f_2(\boldsymbol{x}) & \ldots & f_m(\boldsymbol{x}) \end{pmatrix}^T \overset{law}{=} \begin{pmatrix} f_1(\boldsymbol{y}) & f_2(\boldsymbol{y}) & \ldots & f_m(\boldsymbol{y}) \end{pmatrix}^T.$$

The last results apply to spherical probability distributions, defined as follows (for more details regarding spherical distributions, refer to [21, 29, 4]).

Definition A.11. An $n \times 1$ random vector $\boldsymbol{x}$ is said to have a spherical distribution if for every orthogonal $n \times n$ matrix $\boldsymbol{U}$,

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

Theorem A.12. *(see [30, Theorem 2.1.]) Let $\boldsymbol{x} \overset{law}{=} r\boldsymbol{u}$ be a spherically distributed $n \times 1$ random vector with $\mathbb{P}[\boldsymbol{x} = \boldsymbol{0}] = 0$, where r is independent of $\boldsymbol{u}$ with p.d.f. $h(\cdot)$. Then, the p.d.f. $g(\hat{\boldsymbol{x}})$ of $\boldsymbol{x}$ is given by*

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

B Proof of Theorem 2.5

We prove that $\boldsymbol{y}^* \in S^*$ if and only if $\boldsymbol{B}\boldsymbol{y}^* = \boldsymbol{z}^*$; (2.6) then immediately follows from (2.5). Let $\boldsymbol{y}^* \in \mathbb{R}^d$ be such that $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} \in \mathcal{G}^*$. First, we establish that

$$\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} \in \mathcal{G}^* \text{ if and only if } \boldsymbol{x}^*_\top - \boldsymbol{p}_\top = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{y}^*. \tag{B.1}$$

Suppose that $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} \in \mathcal{G}^*$. Then, using the definition of $\mathcal{G}^*$ (see Definition 2.3) 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

$$\boldsymbol{x}^*_\top - \boldsymbol{p}_\top = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{y}^*. \tag{B.2}$$

Denote by $\boldsymbol{S}$ the $D \times D$ orthogonal matrix $(\boldsymbol{U}\ \boldsymbol{V})$, where $\boldsymbol{V}$ is defined in Assumption 2.2. Using (B.2) 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 $\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p} \in \mathcal{G}^*$. This completes the proof of (B.1).

Now we show that (2.7) and (B.2) 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

$$\begin{pmatrix} \boldsymbol{U}^T \\ \boldsymbol{V}^T \end{pmatrix} (\boldsymbol{x}^*_\top - \boldsymbol{p}_\top) = \begin{pmatrix} \boldsymbol{U}^T \\ \boldsymbol{V}^T \end{pmatrix} \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{A}\boldsymbol{y}^*. \tag{B.3}$$

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 (B.3) 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 $\boldsymbol{V}^T\boldsymbol{U} = \boldsymbol{0}$, which follow from Assumption 2.2. To obtain (B.2) from (2.7), multiply (2.7) by $\boldsymbol{U}$.

C Derivation of the probability density function of $\boldsymbol{w}$

We derive the probability density function of the random vector[15] $\boldsymbol{w}$ defined in (3.3) following a similar line of argument as in [9]: we first derive the distribution of $\|\boldsymbol{w}\|_2^2$ and then show that $\boldsymbol{w}$ follows a spherical distribution, which then allows us to derive the exact distribution of $\boldsymbol{w}$.

Theorem C.1 (Distribution of $\|\boldsymbol{w}\|_2^2$)**.** *Suppose that Assumption 2.2 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. Assume that $\boldsymbol{p}_\top \neq \boldsymbol{x}_\top^*$, where the subscript represents the Euclidean projection on the effective subspace. Let $\boldsymbol{w}$ be defined in* (3.3)*. Then,*

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

where $F(v_1, v_2)$ denotes the F-distribution with degrees of freedom v_1 and v_2.

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 A.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 A.3.

b) Since $\boldsymbol{y}_2^*$ is measurable as a function of $\boldsymbol{B}$ (see proof [9, Lemma A.16]), $\boldsymbol{y}_2^*$ and $\boldsymbol{C}$ must 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 2.7), which is a continuous distribution.

Now, we apply Theorem A.6 to obtain

$$\frac{\|\boldsymbol{w}\|_2^2}{\|\boldsymbol{y}_2^*\|_2^2} = \frac{(\boldsymbol{y}_2^*)^T\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{y}_2^*}{\|\boldsymbol{y}_2^*\|_2^2} \sim \chi^2_{D-d_e}, \tag{C.1}$$

which together with Theorem 2.7 yields

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

where $\chi^2_{D-d_e}$ and $\chi^2_{d-d_e+1}$ are independent[16]. Using the definition of the F-distribution (see Definition A.9), we obtain the desired result. □

[15] For the vector $\boldsymbol{w}$ to be well-defined, we require $d_e < D$ (see Assumption 2.2). If $d_e = D$, then $d = D$; letting $\boldsymbol{Q} = \boldsymbol{I}$ and using $\boldsymbol{z}^* = \boldsymbol{x}^* - \boldsymbol{p}$ in (3.1)–(3.5), it is straightforward to see that $\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] = 1$.

[16] Theorem A.6 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.

Using Theorem C.1, it is straightforward to derive the p.d.f of $\|\boldsymbol{w}\|$.

Theorem C.2 (The p.d.f. of $\|\boldsymbol{w}\|$)**.** *Suppose that Assumption 2.2 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. Assume that $\boldsymbol{p}_\top \neq \boldsymbol{x}^*_\top$, where the subscript represents the Euclidean projection on the effective subspace. The p.d.f. $h(\hat{w})$ of $\|\boldsymbol{w}\|$, with $\boldsymbol{w}$ defined in* (3.3)*, is given by*

$$h(\hat{w}) = \frac{2\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \cdot \frac{\hat{w}^{m-1}}{\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|^m} \left(1 + \frac{\hat{w}^2}{\|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\|^2}\right)^{-(m+n)/2}, \tag{C.3}$$

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

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

$$\|\boldsymbol{w}\| \overset{law}{=} K\sqrt{X}, \tag{C.4}$$

where

$$K = \|\boldsymbol{x}^*_\top - \boldsymbol{p}_\top\| \sqrt{\frac{D - d_e}{d - d_e + 1}}. \tag{C.5}$$

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

$$h(\hat{w}) = \frac{d}{d\hat{w}} \mathbb{P}[\|\boldsymbol{w}\| < \hat{w}] = \frac{d}{d\hat{w}} \mathbb{P}[K\sqrt{X} < \hat{w}] = \frac{d}{d\hat{w}} \mathbb{P}[X < \hat{w}^2/K^2] = \frac{2\hat{w}}{K^2} f(\hat{w}^2/K^2), \tag{C.6}$$

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 (A.2) in (C.6), we obtain the desired result. □

To derive the p.d.f. of $\boldsymbol{w}$ we rely on the fact that $\boldsymbol{w}$ has a spherical distribution (see Definition A.11), as we show next.

Theorem C.3 ($\boldsymbol{w}$ has a spherical distribution)**.** *Suppose that Assumption 2.2 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. Assume that $\boldsymbol{p}_\top \neq \boldsymbol{x}^*_\top$, where the subscript represents the Euclidean projection on the effective subspace. The random vector $\boldsymbol{w}$, defined in* (3.3)*, has a spherical distribution.*

Proof. Our proof is similar to the proof of Lemma A.16 in [9]. 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

$$\boldsymbol{w} \overset{law}{=} \boldsymbol{S}\boldsymbol{w}. \tag{C.7}$$

Using (2.6), 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 C.1. Let $f : \mathbb{R}^{Dd \times 1} \to \mathbb{R}^{(D-d_e)\times 1}$ be a vector-valued function defined as

$$f(\mathrm{vec}[\boldsymbol{C}^T\ \boldsymbol{B}^T]) = \boldsymbol{C}\boldsymbol{B}^T(\boldsymbol{B}\boldsymbol{B}^T)^{-1}\boldsymbol{z}^*, \tag{C.8}$$

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

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

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 A.4, $\mathbb{P}[q = 0] = 0$; this implies that p_i/q is a measurable function for each $i = 1, 2, \dots, D - d_e$ (see [55, Theorem 4.10]).

We have

$$\boldsymbol{w} = f(\mathrm{vec}[\boldsymbol{C}^T\ \boldsymbol{B}^T]) \text{ and } \boldsymbol{S}\boldsymbol{w} = f(\mathrm{vec}[(\boldsymbol{S}\boldsymbol{C})^T\ \boldsymbol{B}^T]). \tag{C.9}$$

From Theorem A.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.10 to conclude that

$$\boldsymbol{w} = f(\mathrm{vec}[\boldsymbol{C}^T\ \boldsymbol{B}^T]) \overset{law}{=} f(\mathrm{vec}[(\boldsymbol{S}\boldsymbol{C})^T\ \boldsymbol{B}^T]) = \boldsymbol{S}\boldsymbol{w}. \qquad \square \tag{C.10}$$

We are now ready to derive the p.d.f. of $\boldsymbol{w}$, and hence prove Theorem 3.3.

Proof of Theorem 3.3: We show that the p.d.f. of $\boldsymbol{w}$ is given by (3.6). The identification with the t-distribution follows from (A.1). 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

$$\mathbb{P}[\boldsymbol{w} = \boldsymbol{0}] = \mathbb{P}[\|\boldsymbol{w}\|^2 = 0] = \mathbb{P}[X = 0], \tag{C.11}$$

where in the last equality we applied Theorem C.1. Since the F-distributed X is a continuous random variable, the last probability in (C.11) is equal to zero.

Since $\mathbb{P}[\boldsymbol{w} = \boldsymbol{0}] = 0$ and $\boldsymbol{w}$ has a spherical distribution (Theorem C.3), Theorem A.12 implies that the p.d.f. $g(\bar{\boldsymbol{w}})$ of $\boldsymbol{w}$ satisfies

$$g(\bar{\boldsymbol{w}}) = \frac{\Gamma(m/2)}{2\pi^{m/2}} h(\|\bar{\boldsymbol{w}}\|)\|\bar{\boldsymbol{w}}\|^{1-m}, \tag{C.12}$$

where $h(\cdot)$ denotes the p.d.f. of $\|\boldsymbol{w}\|$. By substituting (C.3) into (C.12), we obtain the desired result. $\square$

D Proof of Theorem 3.8 and Theorem 3.9

A crucial Lemma is given first.

Lemma D.1. *In the conditions of Theorem 3.8, we have*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \textit{ is successful}] \geq I(\boldsymbol{p}, \Delta), \tag{D.1}$$

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

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

Proof. Theorem 3.7 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 $\boldsymbol{\Sigma} = (\Delta^2/n)\boldsymbol{I}$. According to [29, p. 133],

$$\boldsymbol{w} \overset{law}{=} \frac{\Delta}{s} \begin{pmatrix} Z_1 \\ \vdots \\ Z_m \end{pmatrix}, \tag{D.3}$$

with $s \sim \sqrt{\chi^2_n}$, $m = D - d_e$ and $Z_1, \ldots, Z_m$ i.i.d standard Gaussian random variables. Then, (D.3) yields

$$\begin{aligned}\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] &\geq \mathbb{P}[-\mathbf{1} - \boldsymbol{p}_{d_e+1:D} \leq \boldsymbol{w} \leq \mathbf{1} - \boldsymbol{p}_{d_e+1:D}] \\ &= \mathbb{P}\left[\frac{s}{\Delta}(-1 - p_{d_e+1}) \leq Z_1 \leq \frac{s}{\Delta}(1 - p_{d_e+1}), \ldots, \frac{s}{\Delta}(-1 - p_D) \leq Z_m \leq \frac{s}{\Delta}(1 - p_D)\right], \quad \text{(D.4)}\end{aligned}$$

which can be written as (see [17, p. 1])

$$\int_0^\infty G(\boldsymbol{p}, \Delta, s) h(s) ds, \tag{D.5}$$

where

$$\begin{aligned}G(\boldsymbol{p}, \Delta, s) &= \int_{s(-1-p_{d_e+1})/\Delta}^{s(1-p_{d_e+1})/\Delta} \cdots \int_{s(-1-p_D)/\Delta}^{s(1-p_D)/\Delta} \frac{1}{(2\pi)^{m/2}} e^{-\frac{1}{2}(x_1^2 + \cdots + x_m^2)} dx_1 \ldots dx_m \\ &= \prod_{i=d_e+1}^{D} \frac{1}{\sqrt{2\pi}} \int_{s(-1-p_i)/\Delta}^{s(1-p_i)/\Delta} e^{-x^2/2} dx,\end{aligned} \tag{D.6}$$

and where $h(s)$ is the pdf of s given by

$$h(s) = \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} s^{n-1} e^{-s^2/2}. \tag{D.7}$$

By combining (D.4) – (D.7), we obtain (D.1)–(D.2). □

It is easier to show Theorem 3.9 first, when $\boldsymbol{p} = \mathbf{0}$.

D.1 Proof of Theorem 3.9

The next result is a direct corollary of Lemma D.1 when $\boldsymbol{p} = \mathbf{0}$, allowing us to replace $I(\boldsymbol{p}, \Delta)$ in (D.1) with a new integral $J_{m,n}(\Delta)$ that will be easier to manipulate.

Corollary D.2. *In the conditions and notation of Lemma D.1, let $\boldsymbol{p} = \mathbf{0}$. Then*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \textit{ is successful}] \geq J_{m,n}(\|\boldsymbol{x}^*_\top\|), \tag{D.8}$$

where

$$J_{m,n}(\Delta) := \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \left(\sqrt{\frac{2}{\pi}} \int_0^{s/\Delta} e^{-x^2/2} dx\right)^m s^{n-1} e^{-s^2/2} ds. \tag{D.9}$$

Proof. Let $\boldsymbol{p} = \mathbf{0}$. Then Lemma D.1 implies that $\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] \geq I(\mathbf{0}, \|\boldsymbol{x}^*_\top\|)$, where

$$\begin{aligned}I(\mathbf{0}, \|\boldsymbol{x}^*_\top\|) &= \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \left(\prod_{i=d_e+1}^{D} \frac{1}{\sqrt{2\pi}} \int_{-s/\|\boldsymbol{x}^*_\top\|}^{s/\|\boldsymbol{x}^*_\top\|} e^{-x^2/2} dx\right) s^{n-1} e^{-s^2/2} ds \\ &= \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \left(\sqrt{\frac{2}{\pi}} \int_0^{s/\|\boldsymbol{x}^*_\top\|} e^{-x^2/2} dx\right)^m s^{n-1} e^{-s^2/2} ds \\ &= J_{m,n}(\|\boldsymbol{x}^*_\top\|).\end{aligned} \tag{D.10}$$

□

We need to introduce the following three results on the integral $J_{m,n}(\Delta)$ in (D.9).

Lemma D.3. *The integral $J_{m,n}(\Delta)$ in* (D.9) *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. □

Lemma D.4. *The integral $J_{m,n}(\Delta)$ defined in* (D.9) *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,

$$J_{m,n}(\Delta) \leq \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^{\infty} s^{n-1} e^{-s^2/2} ds = 1.$$

□

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

Theorem D.5. *Let $J_{m,n}(\Delta)$ be the integral defined in* (D.9). *Let n and Δ be fixed and let $r = (n + \Delta^2 - 2)/2$. If $r \neq 0$ then, for large m,*

$$J_{m,n}(\Delta) = \frac{C(n,\Delta)}{(m+1)^{\Delta^2}} \Big((\log(m+1))^r - \frac{r}{2} \log(\log(m+1)) \cdot (\log(m+1))^{r-1} + O((\log(m+1))^{r-1}) \Big), \quad \text{(D.11)}$$

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. The proof of this lemma is similar to the derivations in [56, Section 2, Chapter 2], and is deferred to the end of this appendix. □

Proof of Theorem 3.9 Corollary D.2 implies that

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] \geq I(\mathbf{0}, \|\boldsymbol{x}^*_\top\|) \geq J_{m,n}(\|\boldsymbol{x}^*_\top\|). \quad \text{(D.12)}$$

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 D.3), (D.12) yields

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

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

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

where the inequality follows from Lemma D.4. Thus, (3.12) is satisfied for $\tau_{\mathbf{0}} = J_{m,n}(\sqrt{d_e})$, and (3.15) follows from Theorem D.5. □

D.2 Proof of Theorem 3.8

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 D.2 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 D.6. *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

$$g(z, \Delta, s) = \frac{1}{\sqrt{2\pi}} \int_{s(-1-z)/\Delta}^{s(1-z)/\Delta} e^{-x^2/2} dx, \tag{D.13}$$

and note that

$$I(\boldsymbol{p}, \Delta) = \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \left(\prod_{i=d_e+1}^{D} g(p_i, \Delta, s) \right) s^{n-1} e^{-s^2/2} ds. \tag{D.14}$$

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

$$\begin{aligned} \frac{dg(z, \Delta, s)}{dz} &= e^{\frac{-l(z,\Delta,s)^2}{2}} \frac{d(l(z, \Delta, s))}{dz} - e^{\frac{-l(-z,\Delta,s)^2}{2}} \frac{d(-l(-z, \Delta, s))}{dz} \\ &= e^{\frac{-s^2(1-z)^2}{2\Delta^2}} \left(\frac{-s}{\Delta} \right) - e^{\frac{-s^2(-1-z)^2}{2\Delta^2}} \left(\frac{-s}{\Delta} \right) \\ &= \frac{s}{\Delta} e^{-\frac{s^2}{2\Delta^2}(1+z^2)} \left(e^{-\frac{s^2 z}{\Delta^2}} - e^{\frac{s^2 z}{\Delta^2}} \right). \end{aligned} \tag{D.15}$$

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

$$e^{-\frac{s^2 z}{\Delta^2}} - e^{\frac{s^2 z}{\Delta^2}} = 0, \tag{D.16}$$

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 $z \in [-1, 1]$,

$$g(z, \Delta, s) \geq g(-1, \Delta, s) = \frac{1}{\sqrt{2\pi}} \int_{-l(1,\Delta,s)}^{l(-1,\Delta,s)} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_0^{\frac{2s}{\Delta}} e^{-x^2/2} dx. \tag{D.17}$$

By combining (D.17) with (D.14), we obtain

$$\begin{aligned} I(\boldsymbol{p}, \Delta) &\geq \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \left(\prod_{i=d_e+1}^{D} \frac{1}{\sqrt{2\pi}} \int_0^{\frac{2s}{\Delta}} e^{-x^2/2} dx \right) s^{n-1} e^{-s^2/2} ds \\ &= \frac{1}{2^m} \cdot \frac{1}{2^{\frac{n}{2}-1}\Gamma(\frac{n}{2})} \int_0^\infty \left(\sqrt{\frac{2}{\pi}} \int_0^{\frac{2s}{\Delta}} e^{-x^2/2} dx \right)^m s^{n-1} e^{-s^2/2} ds \\ &= \frac{1}{2^m} J_{m,n}(\Delta/2). \end{aligned} \tag{D.18}$$

□

Proof of Theorem 3.8. Lemma D.1 and Lemma D.6 provide

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] \geq I(\boldsymbol{p}, \Delta) \geq \frac{1}{2^m} J_{m,n}(\Delta/2). \tag{D.19}$$

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 D.3), (D.19) yields

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is successful}] \geq \frac{1}{2^m} J_{m,n}(\sqrt{d_e}) \tag{D.20}$$

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

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

where the inequality follows from Lemma D.4. Thus, (3.12) is satisfied for $\tau = J_{m,n}(\sqrt{d_e})/2^m$, and (3.13) follows from Theorem D.5. □

D.3 Proof of Theorem D.5

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

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

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

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

In [56, Section 2, Chapter 2], Wong derives an asymptotic expansion of a similar integral; our derivations are based on his method.

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

$$J_{m,n}(\Delta) = \frac{\sqrt{\pi}\Delta^n}{\Gamma(\frac{n}{2})} \int_0^\infty e^{Ks(t)^2} s(t)^{n-1} e^{-(m+1)t} dt, \tag{D.21}$$

where $K = 1 - \Delta^2$ and $s(t) = \mathrm{erf}^{-1}(e^{-t})$. 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 (D.21) 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

$$\int_c^\infty e^{Ks(t)^2} s(t)^{n-1} e^{-(m+1)t} dt$$

is exponentially small.

3. Finally, we will derive the asymptotic expansion of

$$\int_0^c e^{Ks(t)^2} s(t)^{n-1} e^{-(m+1)t} dt.$$

Step 1

Lemma D.7. *(see [56, Lemma 1, p. 67]) For small positive t, $s(t) = \operatorname{erf}^{-1}(e^{-t})$ 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 $\operatorname{erf}(s)$ at infinity is given by

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

By writing $1 - e^{-t} = 1 - \operatorname{erf}(s)$ 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

$$\log(t) + O(t) = -s^2 - \log(\sqrt{\pi}) - \log(s) - \frac{1}{2s^2} + O\left(\frac{1}{s^4}\right). \tag{D.22}$$

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

$$s^2 \sim -\log(t), \text{ as } t \to 0^+. \tag{D.23}$$

To obtain higher order approximations, we write

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

and substitute this into (D.22). We have

$$\begin{aligned}\log(t) + O(t) = \log(t) - \epsilon_1(t) - \log(\sqrt{\pi}) - \frac{1}{2}\log(-\log(t)) - \frac{1}{2}\log\left(1 + \frac{\epsilon_1(t)}{-\log(t)}\right) + \\ + O\left(\frac{1}{-\log(t) + \epsilon_1(t)}\right)\end{aligned} \tag{D.24}$$

Note that by (D.23), as $t \to 0^+$

$$\frac{\epsilon_1(t)}{-\log(t)} \to 0. \tag{D.25}$$

By using (D.25) in (D.24), we obtain

$$\epsilon_1(t) = -\frac{1}{2}\log(-\log(t)) - \log(\sqrt{\pi}) + o(1). \tag{D.26}$$

To obtain the following leading terms in the approximation we write

$$s^2(t) = -\log(t) - \frac{1}{2}\log(-\log(t)) - \log(\sqrt{\pi}) + \epsilon_2(t) \tag{D.27}$$

and repeat the above procedure. We substitute (D.27) into (D.22) and after a little manipulation obtain

$$\begin{aligned} O(t) = -\epsilon_2(t) - \frac{1}{2}\log\left(1 - \frac{1}{2}\frac{\log(-\log(t))}{-\log(t)} - \frac{\log(\sqrt{\pi})}{-\log(t)} + \frac{\epsilon_2(t)}{-\log(t)}\right) - \\ - \frac{1}{2}\cdot\frac{1}{-\log(t)}\cdot\frac{1}{1 - \frac{1}{2}\frac{\log(-\log(t))}{-\log(t)} - \frac{\log(\sqrt{\pi})}{-\log(t)} + \frac{\epsilon_2(t)}{-\log(t)}} + O((-\log(t))^2) \end{aligned} \tag{D.28}$$

Using the fact (by (D.26)) 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

$$\epsilon_2(t) = \frac{\log(-\log(t))}{4(-\log(t))} + O\left(\frac{1}{-\log(t)}\right). \tag{D.29}$$

To obtain the following leading terms in the expansion of $\epsilon_2(t)$, we use (D.29) in (D.28) 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

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

Hence,

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

□

Corollary D.8. *Let $l(t) = -\log(t)$. Then, as $t \to 0^+$,*

$$\begin{aligned} e^{Ks(t)^2}s(t)^{n-1} = e^{Kl(t)}\pi^{-K/2}l(t)^{\frac{n-1-K}{2}}\left(1 - \left(\frac{n-1-K}{4}\right)\frac{\log(l(t))}{l(t)} - \frac{\log(e^{K/2}\pi^{\frac{n-1-K}{4}})}{l(t)}\right. \\ \left. + O\left(\frac{\log^2(l(t))}{l(t)^2}\right)\right) \end{aligned} \tag{D.30}$$

Proof. From Theorem D.7 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

$$e^{Ks(t)^2} = e^{Kl(t)} l(t)^{-K/2} \pi^{-K/2} \left(1 + \frac{K \log(l(t))}{4l(t)} - \frac{K \log(e/\sqrt{\pi})}{2l(t)} + O\left(\frac{\log^2(l(t))}{l(t)^2} \right) \right). \quad \text{(D.31)}$$

Similarly, using Theorem D.7 and binomial expansion, for $s(t)^{n-1}$, we have

$$(s(t)^2)^{\frac{n-1}{2}} = l(t)^{\frac{n-1}{2}} \left(1 - \frac{(n-1)\log(l(t))}{4l(t)} - \frac{(n-1)\log(\sqrt{\pi})}{2l(t)} + O\left(\frac{\log^2(l(t))}{l(t)^2} \right) \right) \quad \text{(D.32)}$$

By multiplying the leading terms in (D.31) and (D.32), we obtain the desired result. □

Step 2

Let $0 < c < 1$. We will show that, for large m,

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

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

$$\text{there exists a positive constant } A \text{ such that } s(t) = \mathrm{erf}^{-1}(e^{-t}) \le Ae^{-t} \text{ for all } t \in [c, \infty). \quad \text{(D.33)}$$

Note that (D.33) holds if there exists an $A > 0$ such that $\mathrm{erf}^{-1}(x) \le Ax$ for all $x \in [0, e^{-c}]$. To prove this, we apply the Mean Value Theorem to erf^{-1} over $[0, x]$; by the Mean Value Theorem there exists $y \in (0, x)$ such that

$$\frac{\mathrm{erf}^{-1}(x) - \mathrm{erf}^{-1}(0)}{x - 0} = (\mathrm{erf}^{-1})'(y) \quad \text{(D.34)}$$

Using the following formula for the derivative of the inverse of the error function [1, eq (2.4), p. 192],

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

from (D.34), we obtain

$$\frac{\mathrm{erf}^{-1}(x)}{x} = \frac{\sqrt{\pi}}{2} e^{(\mathrm{erf}^{-1}(y))^2}. \quad \text{(D.35)}$$

Since erf^{-1} is an increasing function and $y < x \le e^{-c}$, (D.35) gives

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

which proves (D.33).

Now, since $s(t)$ is a monotonically decreasing function with $s(\infty) = 0$, we have[17]

$$e^{Ks(t)^2} \le \max\{1, e^{Ks(c)^2}\} \text{ for } t \ge c. \quad \text{(D.36)}$$

Using (D.33) and (D.36), we finally obtain

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

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

Step 3

Let $L(\lambda,\mu,z)$ and $G(\lambda,\mu,z)$ be defined as follows

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

and

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

where $0<c<1$. The expansion of $e^{Ks(t)^2}s(t)^{n-1}$ in Theorem D.8 gives

$$\begin{aligned}\int_0^c e^{Ks(t)^2}s(t)^{n-1}e^{-(m+1)t}dt =&\pi^{-K/2}L\left(1-K,\frac{n-1-K}{2},m+1\right)\\ &-\pi^{-K/2}\left(\frac{n-1-K}{4}\right)G\left(1-K,\frac{n-3-K}{2},m+1\right)\\ &-\pi^{-K/2}\log(e^{K/2}\pi^{\frac{n-1-K}{4}})L\left(1-K,\frac{n-3-K}{2},m+1\right)+\cdots,\end{aligned}\tag{D.37}$$

The following theorem provides the asymptotic expansion for $L(\lambda,\mu,z)$.

Theorem D.9. *(see [56, Theorem 2, p. 70]) 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 rth derivative of the gamma function.

In the following theorem we derive the asymptotic expansion for $G(\lambda,\mu,z)$ based on the proof of [56, Theorem 2, p. 70].

Theorem D.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

$$a_r=-\sum_{i=0}^{r-1}\binom{\mu}{i}\frac{(-1)^i}{r-i}\ \textit{for}\ r=1,2,\ldots.\tag{D.38}$$

Proof. With the substitution $u=zt$, we obtain

$$\begin{aligned}G(\lambda,\mu,z)&=z^{-\lambda}\int_0^{cz}u^{\lambda-1}(\log(z)-\log(u))^{\mu}\log(\log(z)-\log(u))e^{-u}du\\ &=z^{-\lambda}(\log(z))^{\mu}\int_0^{cz}u^{\lambda-1}\left(1-\frac{\log(u)}{\log(z)}\right)^{\mu}\left(\log(\log(z))+\log\left(1-\frac{\log(u)}{\log(z)}\right)\right)e^{-u}du\\ &=z^{-\lambda}(\log(z))^{\mu}(\log(\log(z))G_1+G_2),\end{aligned}\tag{D.39}$$

where

$$G_1 = \int_0^{cz} u^{\lambda-1} \left(1 - \frac{\log(u)}{\log(z)}\right)^{\mu} e^{-u} du$$

and

$$G_2 = \int_0^{cz} u^{\lambda-1} \left(1 - \frac{\log(u)}{\log(z)}\right)^{\mu} \log\left(1 - \frac{\log(u)}{\log(z)}\right) e^{-u} du. \tag{D.40}$$

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,

$$\left(1 - \frac{\log(u)}{\log(z)}\right)^{\mu} = \sum_{r=0}^{N} (-1)^r \binom{\mu}{r} \left(\frac{\log(u)}{\log(z)}\right)^r + R_{1,N}$$

$$\log\left(1 - \frac{\log(u)}{\log(z)}\right) = -\sum_{r=1}^{N} \frac{1}{r} \left(\frac{\log(u)}{\log(z)}\right)^r + R_{2,N},$$

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,

$$\left(1 - \frac{\log(u)}{\log(z)}\right)^{\mu} \log\left(1 - \frac{\log(u)}{\log(z)}\right) = \sum_{r=1}^{2N} a_r \left(\frac{\log(u)}{\log(z)}\right)^r + R_{2N}, \tag{D.41}$$

for all $0 < u < cz$, where a_r's are defined as in (D.38) and

$$|R_{2N}| \leq C_{2N} \frac{|\log(u)|^{2N+1}}{|\log(z)|^{2N+1}}$$

for some fixed $C_{2N} > 0$. By substituting (D.41) in (D.40), 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 [56, 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 [56, eq (2.27), p. 71]; thus, $r_{2N} = O(\log(z)^{-2N-1})$. Hence,

$$G_2 = \sum_{r=1}^{2N} a_r \Gamma^{(r)}(\lambda) (\log(z))^{-r} + O(\log(z)^{-2N-1}). \tag{D.42}$$

In a similar manner, one can show that

$$G_1 = \sum_{r=0}^{N} (-1)^r \binom{\mu}{r} \Gamma^{(r)}(\lambda)(\log(z))^{-r} + O(\log(z)^{-N-1}). \tag{D.43}$$

Combining (D.39), (D.42) and (D.43), we obtain the desired result. □

Conclusions

$$\begin{aligned} J_{m,n}(\Delta) &= \frac{2\Delta^n}{\Gamma(\frac{n}{2})} \int_0^\infty \operatorname{erf}^m(s) s^{n-1} e^{-\Delta^2 s^2} ds \\ &= \frac{\sqrt{\pi}\Delta^n}{\Gamma(\frac{n}{2})} \int_0^\infty e^{Ks(t)^2} s(t)^{n-1} e^{-(m+1)t} dt \\ &= \frac{\sqrt{\pi}\Delta^n}{\Gamma(\frac{n}{2})} \int_0^c e^{Ks(t)^2} s(t)^{n-1} e^{-(m+1)t} dt + O\left(\frac{e^{-c(m+n)}}{m+n}\right). \end{aligned} \tag{D.44}$$

By using Theorem D.9 and Theorem D.10 in (D.37) and substituting $K = 1 - \Delta^2$, we obtain (D.11). 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 $J_{m,1}(1) = 1/(m+1)$.

E Problem set

Table 3 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 [27, 20, 48]. Problems that cannot be solved by BARON are marked with '*'. Problems that will not be solved by KNITRO are marked with '◦'.

We briefly describe the technique we adapted from Wang et al. [54] to generate high-dimensional functions with low effective dimensionality, which was first applied to the above test set in [9]. Let $\bar{g}(\bar{\boldsymbol{x}})$ be any function from Table 3; 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

$$f(\boldsymbol{x}) = g(\boldsymbol{Q}\boldsymbol{x}). \tag{E.1}$$

Note that the first d_e rows of $\boldsymbol{Q}$ now span the effective subspace $\mathcal{T}$ of $f(\boldsymbol{x})$.

For each problem in the test set, we generate three functions f as defined in (E.1), one for each $D = 10$, 100, 1000.

Table 3: The problem set listed in alphabetical order.

Function	Domain	Global minima
1) Beale [20]	$\boldsymbol{x} \in [-4.5, 4.5]^2$	$g(\boldsymbol{x}^*) = 0$
2) *Branin [20]	$x_1 \in [-5, 10]$ $x_2 \in [0, 15]$	$g(\boldsymbol{x}^*) = 0.397887$
3) Brent [27]	$\boldsymbol{x} \in [-10, 10]^2$	$g(\boldsymbol{x}^*) = 0$
4) °Bukin N.6 [48]	$x_1 \in [-15, -5]$ $x_2 \in [-3, 3]$	$g(\boldsymbol{x}^*) = 0$
5) *Easom [20]	$\boldsymbol{x} \in [-100, 100]^2$	$g(\boldsymbol{x}^*) = -1$
6) Goldstein-Price [20]	$\boldsymbol{x} \in [-2, 2]^2$	$g(\boldsymbol{x}^*) = 3$
7) Hartmann 3 [20]	$\boldsymbol{x} \in [0, 1]^3$	$g(\boldsymbol{x}^*) = -3.86278$
8) Hartmann 6 [20]	$\boldsymbol{x} \in [0, 1]^6$	$g(\boldsymbol{x}^*) = -3.32237$
9) *Levy [48]	$\boldsymbol{x} \in [-10, 10]^4$	$g(\boldsymbol{x}^*) = 0$
10) Perm 4, 0.5 [48]	$\boldsymbol{x} \in [-4, 4]^4$	$g(\boldsymbol{x}^*) = 0$
11) Rosenbrock [48]	$\boldsymbol{x} \in [-5, 10]^3$	$g(\boldsymbol{x}^*) = 0$
12) Shekel 5 [48]	$\boldsymbol{x} \in [0, 10]^4$	$g(\boldsymbol{x}^*) = -10.1532$
13) Shekel 7 [48]	$\boldsymbol{x} \in [0, 10]^4$	$g(\boldsymbol{x}^*) = -10.4029$
14) Shekel 10 [48]	$\boldsymbol{x} \in [0, 10]^4$	$g(\boldsymbol{x}^*) = -10.5364$
15) *Shubert [48]	$\boldsymbol{x} \in [-10, 10]^2$	$g(\boldsymbol{x}^*) = -186.7309$
16) Six-hump camel [48]	$x_1 \in [-3, 3]$ $x_2 \in [-2, 2]$	$g(\boldsymbol{x}^*) = -1.0316$
17) Styblinski-Tang [48]	$\boldsymbol{x} \in [-5, 5]^4$	$g(\boldsymbol{x}^*) = -156.664$
18) Trid [48]	$\boldsymbol{x} \in [-25, 25]^5$	$g(\boldsymbol{x}^*) = -30$
19) Zettl [20]	$\boldsymbol{x} \in [-5, 5]^2$	$g(\boldsymbol{x}^*) = -0.00379$