Global optimization using random embeddings

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

24th July 2021

Abstract

We propose a random-subspace algorithmic framework for global optimization of Lipschitz-continuous objectives, and analyse its convergence using novel tools from conic integral geometry. X-REGO randomly projects, in a sequential or simultaneous manner, the high-dimensional original problem into low-dimensional subproblems that can then be solved with any global, or even local, optimization solver. We estimate the probability that the randomly-embedded subproblem shares (approximately) the same global optimum as the original problem. This success probability is then used to show convergence of X-REGO to an approximate global solution of the original problem, under weak assumptions on the problem (having a strictly feasible global solution) and on the solver (guaranteed to find an approximate global solution of the reduced problem with sufficiently high probability). In the particular case of unconstrained objectives with low effective dimension, that only vary over a low-dimensional subspace, we propose an X-REGO variant that explores random subspaces of increasing dimension until finding the effective dimension of the problem, leading to X-REGO globally converging after a finite number of embeddings, proportional to the effective dimension. We show numerically that this variant efficiently finds both the effective dimension and an approximate global minimizer of the original problem.

Keywords: global optimization, random subspaces, conic integral geometry, dimensionality reduction, functions with low effective dimension

1 Introduction

We address the global optimization problem

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

where $f : \mathcal{X} \to \mathbb{R}$ is Lipschitz continuous and possibly non-convex, and where $\mathcal{X}$ is a set with non-empty interior, and possibly unbounded, which thus includes the unconstrained case $\mathcal{X} = \mathbb{R}^D$. We propose a generic algorithmic framework, named X-REGO ($\mathcal{X}$-Random Embeddings for Global Optimization) that (approximately) solves a sequence of realizations of the following randomized reduced problem,

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

[*]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.

[‡]Mathematical Institute, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG, UK; cartis,massart,otemissov@maths.ox.ac.uk

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

where $\boldsymbol{A}$ is a $D \times d$ Gaussian random matrix (see Definition A.1) with $d \ll D$, and where $\boldsymbol{p} \in \mathcal{X}$ may vary between realizations, may be arbitrary/user-defined, and provides additional flexibility that can be exploited algorithmically. The reduced problem (RP$\mathcal{X}$) can be solved by any global, or even local or stochastic, optimization solver.

When a (possibly stochastic) global solver is used in the subproblems, we prove that X-REGO converges, with probability one, to a global ϵ-minimizer of (P) (namely, a feasible point $\boldsymbol{x}$ satisfying $f(\boldsymbol{x}) \leq f^* + \epsilon$ for some accuracy $\epsilon > 0$); we also provide estimates of the corresponding convergence rate. For this, we need to evaluate the *ϵ-success* of the reduced problem (RP$\mathcal{X}$).

Definition 1.1. (RP$\mathcal{X}$) is *ϵ-successful* if there exists $\boldsymbol{y} \in \mathbb{R}^d$ such that $\boldsymbol{A}\boldsymbol{y} + \boldsymbol{p} \in \mathcal{X}$ and $f(\boldsymbol{A}\boldsymbol{y} + \boldsymbol{p}) \leq f^* + \epsilon$, where $\epsilon > 0$ is the desired/user-chosen accuracy tolerance.

Equivalently, this success probability can be rephrased as follows.

What are the chances that a random low-dimensional subspace spanned by the columns of a (rectangular) Gaussian matrix contains a global ϵ-minimizer of (P)*?*

We use crucial tools from conic integral geometry to estimate the probability above. Applications of these bounds to functions with low effective dimensionality are also provided.

1.1 Related work.

Dimensionality reduction is essential to the efficient solution of high-dimensional optimization problems. Sketching techniques reduce the ambient dimension of a given subspace by projecting it randomly into a lower dimensional one while preserving lengths [67]; such techniques have been used successfully for improving the efficiency of linear and nonlinear least squares (local) solvers and of those for more general sums of functions; see for example, [53, 56, 8, 19] and the references therein. Here, we sketch the problem variables/search space in order to reduce its dimension for the specific aim of global optimization; furthermore, our results are not derived using sketching techniques but conic integral geometry ones.

In a huge-scale setting, where full-dimensional vector operations are computationally expensive, Nesterov [49] advocates the use of coordinate descent, a local optimization method that updates successively one of the coordinates of a candidate solution using a coordinate-wise variant of a first-order method, while keeping other coordinates fixed. Coordinate descent methods and their block counterparts have become a method of choice for many large-scale applications, see, e.g., [4, 55, 68] and have been extended to random subspace descent [46, 44] that operates over a succession of random low-dimensional subspaces, not necessarily aligned with coordinate axes. See also [38] for a random proximal subspace descent algorithm, and [35, 40] for higher-order random subspace methods for local nonlinear optimization.

In local derivative-free optimization, several algorithms explore successively one-dimensional [59, 50, 9] and low-dimensional [16] random subspaces. Gratton et al. [36, 37] propose and explore a randomized version of direct search where at each iteration the function is explored along a collection of directions, i.e., one-dimensional half-spaces. Golovin et al. [34] develop convergence rates to a ball of ϵ-minimizers for a variant of randomized direct search for a special class of quasi-convex objectives. Their convergence analysis heavily relies on high-dimensional geometric arguments: they show that sublevel sets contain a sufficiently large ball tangent to the level set, so that at each iteration, with a given probability, sampling the next iterate from a suitable distribution centred at the current iterate decreases the cost.

Unlike the above-mentioned works, our focus here is on the *global* optimization of *generic Lipschitz-continuous objectives*. Stochastic global optimization methods abound, such as simulated annealing [32], random search [58], multistart methods [32], and genetic algorithms [41]. Our proposal here is connected to random search methods, namely, it can be viewed as

a multi-dimensional random search, where a deterministic or stochastic method is applied to the subspace minimization. Recently, random subspace methods have been developed/applied for the global optimization of objectives with special structure, assuming typically, low-effective dimensionality of the objective [66, 10, 11, 43, 15, 18, 54]. These functions only vary over a low-dimensional subspace, and are also called multi-ridge functions [29, 62], functions with active subspaces [21], or functions with functional sparsity when the subspace of variation is aligned with coordinate axes [65]. Assuming the random subspace dimension d (in (RP$\mathcal{X}$)) to be an overestimate of the objective's effective dimension d_e (the dimension of the subspace of variation), these works have proven that one random embedding is sufficient with probability one to solve the original problem (P) in the unconstrained case ($\mathcal{X} = \mathbb{R}^d$) [66, 15] while several random embeddings are required in the constrained case [18]. In particular, in [18], we propose an X-REGO variant that is designed specifically for the bound-constrained optimization of functions with low effective dimensionality. As such it keeps the random subspace dimension d in (RP$\mathcal{X}$) fixed and greater than the effective dimension which is assumed to be known. Here, X-REGO is designed and analysed for a generic objective and a possibly unbounded/unconstrained and nonconvex domain $\mathcal{X}$, and the random subspace dimension d is arbitrary and allowed to vary during the optimization.

Recently, random projections have been successfully applied to highly overparametrized settings, such as in deep neural network training [47, 42] and adversarial attacks in deep learning [14, 63]. Though there is no theoretical guarantee at present that a precise low-dimension subspace exists in these problems, it is a reasonable assumption to make given the high dimensionality of the search space and the supporting numerical evidence. Our approach here investigates the validity of random subspace methods when low effective dimensionality is absent or unknown to the user; we find - both theoretically and numerically - that for large scale problems, such techniques are still beneficial, and furthermore, at least in the unconstrained case, they can naturally adapt and capture such special structures efficiently. We hope that this provides a general theoretical justification to a broader application of such techniques.

The second part of the paper applies the generic X-REGO convergence results and the (RP$\mathcal{X}$) related probabilistic bounds to the case when the objective is unconstrained and has low effective dimension, but the effective dimension d_e is unknown. Related results have been proposed that aim to learn the effective subspace before [29, 24, 62, 27] or during the optimization process [30, 69, 20, 22]; additional costs/evaluations are needed in these approaches. Some apply a principal component analysis (PCA) to the gradient evaluated at a collection of random points [21, 27, 22]. Alternatively, [29, 24, 62] recast the problem into a low-rank matrix recovery problem, and [30] proposes a Bayesian optimization algorithm that sequentially updates a posterior distribution over effective subspaces, and over the objective, using new functions evaluations. Still in the context of Bayesian optimization, Zhang et al. [69] estimate the effective subspace using Sliced Inverse Regression, a supervised dimensionality reduction technique in contrast with the above-mentioned PCA, while Chen et al. [20] extend Sliced Inverse Regression to learn the effective subspace in a semi-supervised way. Instead, our proposed algorithm explores a sequence of random subspaces of increasing dimension until it discovers the effective dimension of the problem. Independently, a similar idea has been recently used in sketching methods for regularized least-squares optimization [45].

Our contributions. We explore the use of random embeddings for the generic global optimization problem (P). Our proposed algorithmic framework, X-REGO, replaces (P) by a sequence of reduced random subproblems (RP$\mathcal{X}$), that are solved (possibly approximately and probabilistically) using any global optimization solver. As such, X-REGO extends block coordinate descent and local random subspace methods to the global setting.

Our convergence analysis for X-REGO crucially relies on a lower bound on the probability

of ϵ-success of (RP$\mathcal{X}$), whose computation, exploiting connections between (RP$\mathcal{X}$) and the field of conic integral geometry, is a key contribution of this paper[1]. Using asymptotic expansions of integrals, we derive interpretable lower bounds in the setting where the random subspace dimension d is fixed and the original dimension D grows to infinity. In the box-constrained case $\mathcal{X} = [-1, 1]^D$, we also compare these bounds with the probability of success of the simplest random search strategy, where a point is sampled in the domain uniformly at random at each iteration. We show that when the point $\boldsymbol{p}$ at which the random subspace is drawn is close enough to a global solution $\boldsymbol{x}^*$ of (P), the random subspace is more likely to intersect a ball of ϵ-minimizer than finding an ϵ-minimizer using random search. Provided that the reduced problem can be solved at a reasonable cost, random subspace methods are thus provably better than random search in some cases; and even more so, numerically.

In the second part of the paper, we address global optimization of functions with low effective dimension, and propose an X-REGO variant that progressively increases the random subspace dimension. Instead of requiring a priori knowledge of the effective dimension of the objective, we show numerically that this variant is able to *learn* the effective dimension of the problem. We also provide convergence results for this variant after a finite number of embeddings, using again our conic integral geometry bounds. Noticeably, these convergence results have no dependency on D. We compare numerically several instances of X-REGO when the reduced problem is solved using the (global and local) KNITRO solver [13]. We also discuss several strategies to choose the parameter $\boldsymbol{p}$ in (RP$\mathcal{X}$).

Paper outline. Section 2 presents the geometry of the problem, and motivates the use of conic integral geometry to estimate the probability of (RP$\mathcal{X}$) being ϵ-successful. Section 3 summarizes key results from conic integral geometry that are used later in the paper. In Section 4, we derive lower bounds on the probability of (RP$\mathcal{X}$) to be ϵ-successful, obtain asymptotic expansions of this probability, and compare the search within random embeddings with random search. Section 5 presents the X-REGO algorithmic framework, and Section 6 the corresponding convergence analysis. Finally, Section 7 proposes a specific instance of X-REGO for global optimization of functions with low effective dimension, with associate convergence results, and Section 8 contains numerical illustrations.

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 $\langle \cdot, \cdot \rangle$, $\| \cdot \|$ (or equivalently $\| \cdot \|_2$) for the usual Euclidean inner product and Euclidean norm, respectively.

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).

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 $x \geq x_0$, $M_1 g(x) \leq |f(x)| \leq M_2 g(x)$.

[1]Note that this is not the first work applying conic integral geometry to optimization, see [1] for an application to the study of phase transitions in random convex optimization problems.

2 Geometric description of the problem

Let $\epsilon > 0$ denote the accuracy to which problem (P) is to be solved, and so let G_ϵ be the set of ϵ-minimizers of (P),

$$G_\epsilon = \{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) \leq f^* + \epsilon\}. \tag{2.1}$$

Note that, by Definition 1.1, the reduced problem (RP$\mathcal{X}$) is ϵ-successful if and only if the intersection of the (affine) subspace $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ and G_ϵ is non-empty:

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] = \mathbb{P}[\boldsymbol{p} + \text{range}(\boldsymbol{A}) \cap G_\epsilon \neq \varnothing]. \tag{2.2}$$

To further characterize this probability, let us now introduce the following assumptions.

Assumption LipC (Lipschitz continuity of f)**.** The objective function $f : \mathcal{X} \to \mathbb{R}$ is Lipschitz continuous with constant L, i.e., there holds $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$.

Assumption FeasBall (Existence of a ball of ϵ-minimizers)**.** There exists a global minimizer $\boldsymbol{x^*}$ of (P) that satisfies $B_{\epsilon/L}(\boldsymbol{x}^*) \subset \mathcal{X}$, where $B_{\epsilon/L}(\boldsymbol{x}^*)$ is the D-dimensional Euclidean ball of radius ϵ/L and centered at $\boldsymbol{x}^*$, where L is the Lipschitz constant of f and $\epsilon > 0$ is the desired accuracy tolerance.

We then have the following result.

Proposition 2.1. *Let Assumption LipC hold. Let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix, ϵ a positive accuracy tolerance and $\boldsymbol{x}^*$ any global minimizer of* (P) *satisfying Assumption FeasBall. Let $\boldsymbol{p} \in \mathcal{X}$ be a given vector. Then,*

$$\mathbb{P}[(\text{RP}\mathcal{X}) \textit{ is } \epsilon\textit{-successful}] \geq \mathbb{P}[\boldsymbol{p} + \text{range}(\boldsymbol{A}) \cap B_{\epsilon/L}(\boldsymbol{x}^*) \neq \varnothing]. \tag{2.3}$$

Proof. Let $\boldsymbol{x}^*$ be a global minimizer of f in $\mathcal{X}$ satisfying Assumption FeasBall, and let $\boldsymbol{x} \in B_{\epsilon/L}(\boldsymbol{x}^*)$. Then, $\boldsymbol{x} \in G_\epsilon$ due to the Lipschitz continuity property of f, namely

$$|f(\boldsymbol{x}) - f(\boldsymbol{x}^*)| \leq L\|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \leq L\frac{\epsilon}{L} = \epsilon. \tag{2.4}$$

The result follows then simply from (2.2). □

In the case of non-unique solutions, each global minimizer $\boldsymbol{x}^*$ of (P) satisfying Assumption FeasBall provides a different lower bound in Proposition 2.1. If all the balls $B_{\epsilon/L}(\boldsymbol{x}^*)$ associated with different global minimizers are disjoint, the probability of ϵ-success of (RP$\mathcal{X}$) is lower bounded by the sum, over each $\boldsymbol{x}^*$ satisfying Assumption FeasBall, of the probability $\mathbb{P}[\boldsymbol{p} + \text{range}(\boldsymbol{A}) \cap B_{\epsilon/L}(\boldsymbol{x}^*) \neq \varnothing]$. In this paper, we estimate the latter probability for an arbitrary $\boldsymbol{x}^*$; this is a worst-case bound in the sense that it clearly underestimates the chance of subproblem success (for a(ny) $\boldsymbol{x}^*$) in the presence of multiple global minimizers of (P).

Given $\boldsymbol{x}^*$ satisfying Assumption FeasBall, let us assume that $\boldsymbol{p} \notin B_{\epsilon/L}(\boldsymbol{x}^*)$ (otherwise, the reduced problem (RP$\mathcal{X}$) is always ϵ-successful, which can be seen by simply taking $\boldsymbol{y} = \boldsymbol{0}$). To estimate the right-hand side of (2.3), we first construct a set $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ containing the rays connecting $\boldsymbol{p}$ with points in $B_{\epsilon/L}(\boldsymbol{x}^*)$,

$$C_{\boldsymbol{p}}(\boldsymbol{x}^*) = \{\boldsymbol{p} + \theta(\boldsymbol{x} - \boldsymbol{p}) : \theta \geq 0, \boldsymbol{x} \in B_{\epsilon/L}(\boldsymbol{x}^*)\} \text{ for } \boldsymbol{p} \notin B_{\epsilon/L}(\boldsymbol{x}^*). \tag{2.5}$$

Note that $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ is a convex cone that has been translated by $\boldsymbol{p}$ (see Figure 1). We can easily verify this fact by recalling the definition of a convex cone.

Definition 2.2. A convex set C is called a convex cone if for every $\boldsymbol{c} \in C$ and any non-negative scalar ρ, $\rho\boldsymbol{c} \in C$.

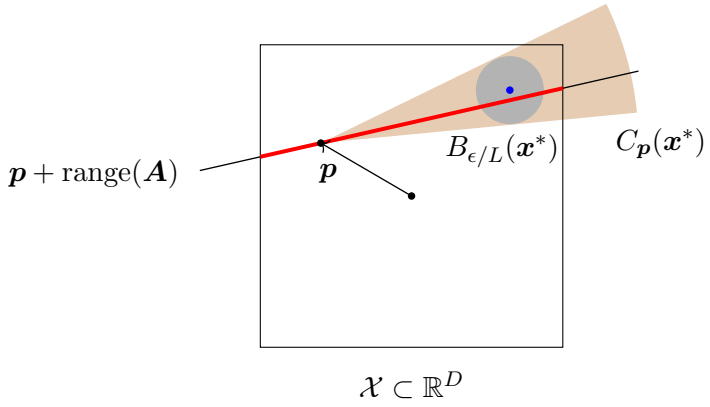

Figure 1: Abstract illustration of the embedding of an affine d-dimensional subspace $\boldsymbol{p}+\text{range}(\boldsymbol{A})$ into $\mathbb{R}^D$, in the case $\mathcal{X} = [-1, 1]^D$. The red line represents the set of solutions along $\boldsymbol{p}+\text{range}(\boldsymbol{A})$ that are contained in $\mathcal{X}$ and the blue dot represents a global minimizer $\boldsymbol{x}^*$ of (P). (RP$\mathcal{X}$) is ϵ-successful when the red line intersects $B_{\epsilon/L}(\boldsymbol{x}^*)$. We construct a cone $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ in such a way that the following condition holds: $\boldsymbol{p}+\text{range}(\boldsymbol{A})$ intersects $B_{\epsilon/L}(\boldsymbol{x}^*)$ if and only if $\boldsymbol{p}+\text{range}(\boldsymbol{A})$ and $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ share a ray.

Remark 2.3. Note that, according to Definition 2.2, a d-dimensional linear subspace in $\mathbb{R}^D$ is a cone. Hence, range$(\boldsymbol{A})$ is a cone.

The next result indicates that, based on (2.3) and the definition of $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$, we can rewrite the right-hand side of (2.3) as

$$\mathbb{P}[\boldsymbol{p} + \text{range}(\boldsymbol{A}) \cap B_{\epsilon/L}(\boldsymbol{x}^*) \neq \varnothing] = \mathbb{P}[\boldsymbol{p} + \text{range}(\boldsymbol{A}) \cap C_{\boldsymbol{p}}(\boldsymbol{x}^*) \neq \{\boldsymbol{p}\}] \tag{2.6}$$

— the probability of the event that translated cones $\boldsymbol{p} + \text{range}(\boldsymbol{A})$ and $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ share a ray. It turns out that this probability has a quantifiable expression based on conic integral geometry, where a broad concern is the quantification/estimation of probabilities of a random cone (e.g., $\boldsymbol{p} + \text{range}(\boldsymbol{A})$) and a fixed cone (e.g., $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$) sharing a ray. We then present in Section 3 key tools from conic integral geometry to help us estimate the probability of ϵ-success of (RP$\mathcal{X}$).

Theorem 2.4. *Let Assumption LipC hold. Let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix, ϵ a positive accuracy tolerance and $\boldsymbol{x}^*$ any global minimizer of* (P) *satisfying Assumption FeasBall. Let $\boldsymbol{p} \in \mathcal{X} \setminus G_\epsilon$ be a given vector and let $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ be defined in* (2.5)*. Then,*

$$\mathbb{P}[(\text{RP}\mathcal{X}) \textit{ is } \epsilon\textit{-successful}] \geq \mathbb{P}[\boldsymbol{p} + \text{range}(\boldsymbol{A}) \cap C_{\boldsymbol{p}}(\boldsymbol{x}^*) \neq \{\boldsymbol{p}\}]. \tag{2.7}$$

Proof. From Proposition 2.1, we have

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

The result follows from the fact that the event $\{\boldsymbol{p} + \text{range}(\boldsymbol{A}) \cap C_{\boldsymbol{p}}(\boldsymbol{x}^*) \neq \{\boldsymbol{p}\}\}$ is a subset of the event $\{\boldsymbol{p} + \text{range}(\boldsymbol{A}) \cap B_{\epsilon/L}(\boldsymbol{x}^*) \neq \varnothing\}$. We prove this fact below.

Suppose that the event $\{\boldsymbol{p} + \text{range}(\boldsymbol{A}) \cap C_{\boldsymbol{p}}(\boldsymbol{x}^*) \neq \{\boldsymbol{p}\}\}$ occurs. Then, there exists a point $\boldsymbol{x}' \neq \boldsymbol{p}$ in $\boldsymbol{p} + \text{range}(\boldsymbol{A}) \cap C_{\boldsymbol{p}}(\boldsymbol{x}^*)$. Define $R = \{\boldsymbol{p} + \theta(\boldsymbol{x}' - \boldsymbol{p}) : \theta \geq 0\}$ and note that $R \subset \boldsymbol{p} + \text{range}(\boldsymbol{A})$. Now, since $\boldsymbol{x}' \in C_{\boldsymbol{p}}(\boldsymbol{x}^*)$, by definition of $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ there exists $\tilde{\boldsymbol{x}} \in B_{\epsilon/L}(\boldsymbol{x}^*)$ and $\tilde{\theta} > 0$ such that $\boldsymbol{x}' = \boldsymbol{p} + \tilde{\theta}(\tilde{\boldsymbol{x}} - \boldsymbol{p})$. We express $\tilde{\boldsymbol{x}}$ in terms of $\boldsymbol{x}'$: $\tilde{\boldsymbol{x}} = \boldsymbol{p} + \theta'(\boldsymbol{x}' - \boldsymbol{p})$, where $\theta' = 1/\tilde{\theta} > 0$. By definition of R, $\tilde{\boldsymbol{x}} \in R$ and, thus, $\tilde{\boldsymbol{x}}$ also lies in $\boldsymbol{p} + \text{range}(\boldsymbol{A})$. This proves that the set $\{\boldsymbol{p} + \text{range}(\boldsymbol{A}) \cap B_{\epsilon/L}(\boldsymbol{x}^*)\}$ is non-empty. □

3 A snapshot of conic integral geometry

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

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

The answer to this question is given by the conic kinematic formula [57].

Theorem 3.1 (Conic kinematic formula)**.** *Let C and F be closed convex cones in $\mathbb{R}^D$ such that at most one of them is a linear subspace. Let $\boldsymbol{Q}$ be a $D \times D$ random orthogonal matrix drawn uniformly from the set of all $D \times D$ real orthogonal matrices. Then,*

$$\mathbb{P}[\boldsymbol{Q}F \cap C \neq \{\mathbf{0}\}] = \sum_{k=0}^{D}(1 + (-1)^{k+1}) \sum_{j=k}^{D} v_k(C)v_{D+k-j}(F), \tag{3.1}$$

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

Proof. A proof can be found in [57, p. 261]. □

We plan to use the conic kinematic formula to estimate (2.6). This formula expresses the probability of the intersection of the two cones in terms of quantities known as conic intrinsic volumes. It is thus important to understand the conic intrinsic volumes and ways to compute them.

3.1 Conic intrinsic volumes

Conic intrinsic volumes are commonly defined through the spherical Steiner formula (see [2]), which we do not define here as it is beyond the scope of this work/not needed here. Instead, we will familiarise ourselves with the conic intrinsic volumes through their properties and specific examples. This is a short introductory review of conic intrinsic volumes; for more details, an interested reader is directed to [2, 3, 1, 48, 57] and the references therein.

For a closed convex cone C in $\mathbb{R}^D$, there are exactly $D + 1$ conic intrinsic volumes: $v_0(C)$, $v_1(C)$,..., $v_D(C)$. Conic intrinsic volumes have useful properties, some of which are summarized below. Given a closed convex cone $C \subseteq \mathbb{R}^D$, we have (see [3, Fact 5.5]):

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

$$\sum_{k=0}^{D} v_k(C) = 1 \text{ and } v_k(C) \geq 0 \text{ for } k = 0, 1, \ldots, D. \tag{3.2}$$

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

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

$$v_k(\boldsymbol{Q}C) = v_k(C). \tag{3.3}$$

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

$$\sum_{\substack{k=0 \\ k \text{ even}}}^{D} v_k(C) = \sum_{\substack{k=1 \\ k \text{ odd}}}^{D} v_k(C) = \frac{1}{2}. \tag{3.4}$$

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

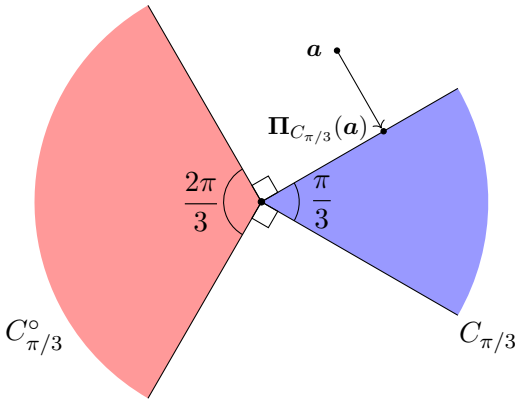

Figure 2: A depiction of the two-dimensional polyhedral cone $C_{\pi/3}$ in Example 3.4. The projection $\boldsymbol{\Pi}_{C_{\pi/3}}(\boldsymbol{a})$ of $\boldsymbol{a}$ onto $C_{\pi/3}$ falls onto the one-dimensional face of the cone.

Remark 3.2. Conic intrinsic volumes can be viewed as 'cousins' of the more familiar Euclidean intrinsic volumes. For a compact convex set K living in $\mathbb{R}^D$, Euclidean intrinsic volumes $v_0^E(K)$, $v_{D-1}^E(K)$ and $v_D^E(K)$ have familiar geometric interpretations: $v_0^E(K)$ — Euler characteristic, $2v_{D-1}^E(K)$ — surface area and $v_D^E(K)$ is the usual volume.

Remark 3.3. Conic intrinsic volumes can also be understood using polyhedral cones — cones that can be generated by intersecting a finite number of halfspaces. If C is a polyhedral cone in $\mathbb{R}^D$, then the kth intrinsic volume of C is defined as follows (see [3, Definition 5.1])

$$v_k(C) := \mathbb{P}[\boldsymbol{\Pi}_C(\boldsymbol{a}) \text{ lies in the relative interior}^2 \text{of a } k\text{-dimensional face of } C]. \tag{3.5}$$

Here, $\boldsymbol{a}$ denotes the standard Gaussian vector[3] in $\mathbb{R}^D$ and $\boldsymbol{\Pi}_Y(\boldsymbol{x}) := \arg\min_{\boldsymbol{y}}\{\|\boldsymbol{x}-\boldsymbol{y}\| : \boldsymbol{y} \in Y\}$ denotes the Euclidean/orthogonal projection of $\boldsymbol{x}$ onto the set Y, namely the vector in Y that is the closest to $\boldsymbol{x}$.

Example 3.4. Let us consider a simple a two-dimensional polyhedral cone $C_{\pi/3}$ illustrated in Figure 2 and let us calculate $v_0(C_{\pi/3})$, $v_1(C_{\pi/3})$ and $v_2(C_{\pi/3})$ using (3.5).

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

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

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

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

To calculate $v_1(C_{\pi/3})$, we simply use (3.2) to obtain

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

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

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

Example 3.5 (Linear subspace)**.** The kth intrinsic volume of a d-dimensional linear subspace $\mathcal{L}_d$ in $\mathbb{R}^D$ is given by

$$v_k(\mathcal{L}_d) = \begin{cases} 1 & \text{if } k = d, \\ 0 & \text{otherwise.} \end{cases} \tag{3.6}$$

We already mentioned in Remark 2.3 that a d-dimensional linear subspace $\mathcal{L}_d$ is a cone. In fact, $\mathcal{L}_d$ is a polyhedral cone which has only one (d-dimensional) face. Therefore, the projection of any vector in $\mathbb{R}^D$ onto $\mathcal{L}_d$ will always lie on its (only) d-dimensional face. Hence, (3.6) follows from (3.5).

Example 3.6 (Circular cone)**.** A circular cone is another important example; they have a number of applications in convex optimization (see, e.g., [7, Section 3] and [12, Section 4]). The circular cone of angle α in $\mathbb{R}^D$ is denoted by $\mathrm{Circ}_D(\alpha)$ and is defined as

$$\mathrm{Circ}_D(\alpha) := \{\boldsymbol{x} \in \mathbb{R}^D : x_1 \geq \|\boldsymbol{x}\| \cos(\alpha)\} \text{ for } 0 \leq \alpha \leq \pi/2. \tag{3.7}$$

The circular cone can be viewed as a collection of rays connecting the origin and some D-dimensional ball which does not contain the origin in its interior. The intrinsic volumes of $\mathrm{Circ}_D(\alpha)$ are given by the formulae (see [3, Appendix D.1]):

$$v_k(\mathrm{Circ}_D(\alpha)) = \frac{1}{2}\binom{(D-2)/2}{(k-1)/2} \sin^{k-1}(\alpha)\cos^{D-k-1}(\alpha) \tag{3.8}$$

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

$$\binom{i}{j} = \frac{\Gamma(i+1)}{\Gamma(j+1)\Gamma(i-j+1)}. \tag{3.9}$$

The 0th and Dth intrinsic volumes of the circular cone are given by (see [1, Ex. 4.4.8]):

$$v_0(\mathrm{Circ}_D(\alpha)) = \frac{D-1}{2}\binom{(D-2)/2}{-1/2} \int_0^{\pi/2-\alpha} \sin^{D-2}(x)dx, \tag{3.10}$$

$$v_D(\mathrm{Circ}_D(\alpha)) = \frac{D-1}{2}\binom{(D-2)/2}{(D-1)/2} \int_0^{\alpha} \sin^{D-2}(x)dx. \tag{3.11}$$

The following property of circular cones will be needed later.

Lemma 3.7. *Let* $\mathrm{Circ}_D(\alpha)$ *and* $\mathrm{Circ}_D(\beta)$ *be two circular cones with* $0 \leq \alpha \leq \beta \leq \pi/2$*. Then,* $\mathrm{Circ}_D(\alpha) \subseteq \mathrm{Circ}_D(\beta)$*.*

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

3.2 The Crofton formula

We now present a useful corollary of the conic kinematic formula. If one of the cones in Theorem 3.1 is given by a linear subspace then the conic kinematic formula reduces to the Crofton formula.

Corollary 3.8 (Crofton formula)**.** *Let* C *be a closed convex cone in* $\mathbb{R}^D$ *and* $\mathcal{L}_d$ *be a* d*-dimensional linear subspace. Let* $\boldsymbol{Q}$ *be a* $D \times D$ *random orthogonal matrix drawn uniformly from the set of all* $D \times D$ *real orthogonal matrices. We have*

$$\mathbb{P}[\boldsymbol{Q}\mathcal{L}_d \cap C \neq \{\boldsymbol{0}\}] = 2h_{D-d+1}, \tag{3.12}$$

with

$$h_{D-d+1} := \begin{cases} v_{D-d+1}(C) + v_{D-d+3}(C) + \cdots + v_D(C) & \text{if } d \text{ is odd,} \\ v_{D-d+1}(C) + v_{D-d+3}(C) + \cdots + v_{D-1}(C) & \text{if } d \text{ is even.} \end{cases} \tag{3.13}$$

The Crofton formula is easily derived from (3.1) using the fact that the kth intrinsic volume of a linear subspace $\mathcal{L}_d$ is 1 if $d = k$ and 0 otherwise. The Crofton formula will be essential in estimating the probability of ϵ-success of (RP$\mathcal{X}$).

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

Building on the tools developed in the last section, we can estimate the right-hand side of (2.7) in Theorem 2.4, and thereby obtain bounds on the probability of ϵ-success of (RP$\mathcal{X}$).

Note that if $\boldsymbol{p} \notin B_{\epsilon/L}(\boldsymbol{x}^*)$, then $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ defined in (2.5) is a circular cone $\mathrm{Circ}_D(\alpha^*_{\boldsymbol{p}})$ with $\alpha^*_{\boldsymbol{p}} = \arcsin(\epsilon/(L\|\boldsymbol{x}^* - \boldsymbol{p}\|))$ that has been rotated and then translated by $\boldsymbol{p}$, see (3.7). Therefore, the intersection $\boldsymbol{p} + \mathrm{range}(\boldsymbol{A}) \cap C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ in (2.7) is that of a random d-dimensional linear subspace and the rotated circular cone both translated by $\boldsymbol{p}$. We can translate these 'cones' back to the origin and then, using the Crofton formula, evaluate the right-hand side of (2.7) exactly since the expressions for the conic intrinsic volumes of the circular cone $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ are known (see (3.8), (3.10) and (3.11)). The Crofton formula and the right-hand side of (2.7) only differ in the formulation of a random linear subspace: in the former, a random linear subspace is given as $\boldsymbol{Q}\mathcal{L}_d$, whereas in (2.7) it is represented by $\mathrm{range}(\boldsymbol{A})$. The following theorem states that these two representations are equivalent.

Theorem 4.1. *Let $\boldsymbol{A} \in \mathbb{R}^{D\times d}$ be a Gaussian matrix. Let $\boldsymbol{Q}$ be a $D \times D$ random orthogonal matrix drawn uniformly from the set of all $D \times D$ real orthogonal matrices and let $\mathcal{L}_d$ be a d-dimensional linear subspace in $\mathbb{R}^D$. Then,*

$$\mathrm{range}(\boldsymbol{A}) \overset{law}{=} \boldsymbol{Q}\mathcal{L}_d. \tag{4.1}$$

Proof. See proof of [33, Theorem 1.2]. □

The transformation of (2.7) into a form suitable for the application of Crofton formula is given in the following corollary.

Corollary 4.2. *Let Assumption LipC hold. Let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix, $\boldsymbol{Q}$ be a $D \times D$ random orthogonal matrix drawn uniformly from the set of all $D\times D$ real orthogonal matrices and $\mathcal{L}_d$ be a d-dimensional linear subspace in $\mathbb{R}^D$. Let $\epsilon > 0$ an accuracy tolerance and let $\boldsymbol{p} \in \mathcal{X} \setminus G_\epsilon$ be a given vector. Let $\mathrm{Circ}_D(\alpha^*_{\boldsymbol{p}})$ be the circular cone with $\alpha^*_{\boldsymbol{p}} = \arcsin(\epsilon/(L\|\boldsymbol{x}^* - \boldsymbol{p}\|))$, where $\boldsymbol{x}^*$ is any global minimizer of (P) satisfying Assumption FeasBall. Then,*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] \geq \mathbb{P}[\boldsymbol{Q}\mathcal{L}_d \cap \mathrm{Circ}_D(\alpha^*_{\boldsymbol{p}}) \neq \{\boldsymbol{0}\}]. \tag{4.2}$$

Proof. As mentioned earlier, by definition, $C_{\boldsymbol{p}}(\boldsymbol{x}^*)$ is the rotated and translated (by $\boldsymbol{p}$) circular cone $\mathrm{Circ}_D(\alpha^*_{\boldsymbol{p}})$. That is, there exists a $D \times D$ orthogonal matrix $\boldsymbol{S}$ such that $C_{\boldsymbol{p}}(\boldsymbol{x}^*) = \boldsymbol{p} + \boldsymbol{S}\,\mathrm{Circ}_D(\alpha^*_{\boldsymbol{p}})$. Then, Theorem 2.4 implies

$$\begin{aligned} \mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] &\geq \mathbb{P}[\boldsymbol{p} + \mathrm{range}(\boldsymbol{A}) \cap \boldsymbol{p} + \boldsymbol{S}\,\mathrm{Circ}_D(\alpha^*_{\boldsymbol{p}}) \neq \{\boldsymbol{p}\}] \\ &= \mathbb{P}[\mathrm{range}(\boldsymbol{A}) \cap \boldsymbol{S}\,\mathrm{Circ}_D(\alpha^*_{\boldsymbol{p}}) \neq \{\boldsymbol{0}\}] \\ &= \mathbb{P}[\boldsymbol{S}^T\,\mathrm{range}(\boldsymbol{A}) \cap \mathrm{Circ}_D(\alpha^*_{\boldsymbol{p}}) \neq \{\boldsymbol{0}\}] \\ &= \mathbb{P}[\mathrm{range}(\boldsymbol{A}) \cap \mathrm{Circ}_D(\alpha^*_{\boldsymbol{p}}) \neq \{\boldsymbol{0}\}] \\ &= \mathbb{P}[\boldsymbol{Q}\mathcal{L}_d \cap \mathrm{Circ}_D(\alpha^*_{\boldsymbol{p}}) \neq \{\boldsymbol{0}\}], \end{aligned} \tag{4.3}$$

where the penultimate equality follows from the orthogonal invariance of Gaussian matrices and where the last equality follows from Theorem 4.1. □

Corollary 4.2 now allows us to use the Crofton formula to quantify the lower bound in (4.2). In the next theorem, we derive our first lower bound, that is dependent on the location of $\boldsymbol{p}$ in $\mathcal{X}$. In particular, note that $\boldsymbol{p}$ is assumed to be at a distance at least ϵ/L from $\boldsymbol{x}^*$.

Theorem 4.3 (A lower bound on the success probability)**.** *Let Assumption LipC hold, let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix and $\epsilon > 0$, an accuracy tolerance. Let $\boldsymbol{p} \in \mathcal{X} \setminus G_\epsilon$ be a given vector and let $r_{\boldsymbol{p}} := \epsilon/(L\|\boldsymbol{x}^* - \boldsymbol{p}\|)$, where $\boldsymbol{x}^*$ is any global minimizer of* (P) *that satisfies Assumption FeasBall. Then,*

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] \geq \tau = \tau(r_{\boldsymbol{p}}, d, D), \tag{4.4}$$

where the function $\tau(r, d, D)$ for $0 < r < 1$ and $1 \leq d < D$ is defined as

$$\tau(r,d,D) := \begin{cases} (D-1) \cdot \binom{\frac{D-2}{2}}{\frac{D-1}{2}} \displaystyle\int_0^{\arcsin(r)} \sin^{D-2}(x)dx & \text{if } d = 1, \\ \binom{\frac{D-2}{2}}{\frac{D-d}{2}} r^{D-d}(1-r^2)^{\frac{d-2}{2}} & \text{if } 1 < d < D. \end{cases} \tag{4.5}$$

Here, $\binom{i}{j}$ denotes the general binomial coefficient defined in (3.9)*.*

Proof. Let $\alpha^*_{\boldsymbol{p}} = \arcsin(r_{\boldsymbol{p}})$ and let C denote $\mathrm{Circ}_D(\alpha^*_{\boldsymbol{p}})$ for notational convenience. First, note that by (3.8) and (3.11), $\tau(r, d, D) = 2v_{D-d+1}(\mathrm{Circ}_D(\arcsin(r)))$. Thus, all we need to show is that $\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}]$ is lower bounded by $2v_{D-d+1}(C)$.

By (4.2) and the Crofton formula (3.12), we have

$$\begin{aligned} \mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] &\geq \begin{cases} 2(v_{D-d+1}(C) + v_{D-d+3}(C) + \cdots + v_D(C)) & \text{if } d \text{ is odd}, \\ 2(v_{D-d+1}(C) + v_{D-d+3}(C) + \cdots + v_{D-1}(C)) & \text{if } d \text{ is even} \end{cases} \\ &\geq 2v_{D-d+1}(C), \end{aligned} \tag{4.6}$$

where the inequality follows from the fact that $v_k(C)$'s are all nonnegative (see (3.2)). □

Let us explain why we choose to bound the ϵ-success of $(\mathrm{RP}\mathcal{X})$ in (4.6) by a multiple of $v_{D-d+1}(C)$ in particular, whereas we could have chosen any other intrinsic volume or the entire sum of these volumes. Our reason for such a choice for the lower bound is underpinned by the following observation: using the formulae (3.8) and (3.11) for the intrinsic volumes, one can verify that $v_{D-d+i}(C)/v_{D-d+1}(C) = \mathcal{O}(D^{(1-i)/2})$ for $i = 1, 2, \ldots, d$ as $D \to \infty$ with other parameters kept fixed[4]. Hence,

$$v_{D-d+1}(C) + v_{D-d+3}(C) + \cdots = v_{D-d+1}(C) \cdot (1 + \mathcal{O}(1/D)).$$

Therefore, approximating the sum by its leading term $v_{D-d+1}(C)$ is reasonable for large values of D.

Given a global minimizer $\boldsymbol{x}^*$ of (P) that satisfies Assumption FeasBall and a positive constant $R_{\max}$, the following result provides a lower bound on the probability of ϵ-success of $(\mathrm{RP}\mathcal{X})$ that holds for all $\boldsymbol{p} \in \mathcal{X}$ satisfying $\|\boldsymbol{x}^* - \boldsymbol{p}\| \leq R_{\max} < \infty$. Note that, in contrast with the last theorem, this result holds for $\boldsymbol{p}$ arbitrarily close to $\boldsymbol{x}^*$; as such, it will be crucial to the convergence of our algorithmic proposals in Section 6. Note that there are natural ways to choose $R_{\max}$ in some cases:

- If a sequence of reduced problems $(\mathrm{RP}\mathcal{X})$ is being considered such that the random subspaces are drawn at the same $\boldsymbol{p} \in \mathcal{X}$, on can simply take $R_{\max} = \|\boldsymbol{x}^* - \boldsymbol{p}\|$.

[4]The term $v_{D-d+1}(C)$ is dominant also in the scenario when $\|\boldsymbol{x}^* - \boldsymbol{p}\| \to \infty$ as $D \to \infty$ with other parameters fixed. In this case, $v_{D-d+i}(C)/v_{D-d+1}(C) = \mathcal{O}((r_{\boldsymbol{p}}/\sqrt{D})^{i-1})$ for $i = 1, 2, \ldots, d$ as $D \to \infty$.

- If the sequence of reduced problems (RP$\mathcal{X}$) corresponds to a bounded parameter sequence $\{\boldsymbol{p}^0, \boldsymbol{p}^1, \dots\}$, one can choose $R_{\max}$ to be the (finite) supremum over the sequence $\{\|\boldsymbol{x}^* - \boldsymbol{p}^i\|\}$ for $i \geq 0$.
- If $\mathcal{X}$ is bounded, since $\boldsymbol{p} \in \mathcal{X}$ and $\boldsymbol{x}^* \in \mathcal{X}$, one can simply choose $R_{\max}$ to be the diameter of $\mathcal{X}$.

Note that when $\mathcal{X}$ is not bounded, it is in general difficult to derive a uniform lower bound on the probability of ϵ-success of (RP$\mathcal{X}$) that is valid for all $\boldsymbol{p} \in \mathcal{X}$ (taking $\boldsymbol{p} \to \boldsymbol{\infty}$ will make the lower bound go to zero). The above list provides two examples of rules for selecting $\boldsymbol{p}$ that guarantee that the result below holds even in the case $\mathcal{X}$ bounded. Other examples are given in Section 5.

Theorem 4.4 (A uniform lower bound on the success probability)**.** *Suppose that Assumption LipC holds. Let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix, ϵ a positive accuracy tolerance, $\boldsymbol{x}^*$ a global minimizer of* (P) *that satisfies Assumption FeasBall. For all $\boldsymbol{p} \in \mathcal{X}$ satisfying $\|\boldsymbol{p} - \boldsymbol{x}^*\| < R_{\max}$ for some suitably chosen constant $R_{\max}$, we have*

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] \geq \tau = \tau(r_{min}, d, D), \tag{4.7}$$

where $\tau(\cdot,\cdot,\cdot)$ is defined in (4.5) *and $r_{min} := \epsilon/(LR_{\max})$.*

Proof. Let $\boldsymbol{x}^*$ be a global minimizer that satisfies Assumption FeasBall, let $r_{\boldsymbol{p}}$ be defined in Theorem 4.3 and let $\alpha^*_{\boldsymbol{p}} = \arcsin(r_{\boldsymbol{p}})$. We consider the two cases $\boldsymbol{p} \in \mathcal{X} \setminus G_\epsilon$ and $\boldsymbol{p} \in G_\epsilon$ separately.

First, let $\boldsymbol{p}$ be any point in $\mathcal{X} \setminus G_\epsilon$. Then,

$$\begin{aligned} r_{\boldsymbol{p}} &= \frac{\epsilon}{L\|\boldsymbol{p} - \boldsymbol{x}^*\|} \geq r_{min}, \\ \alpha^*_{\boldsymbol{p}} &\geq \arcsin(r_{min}) := \alpha^*_{min}. \end{aligned} \tag{4.8}$$

Now, define $C_{min} := \text{Circ}_D(\alpha^*_{min})$. By (4.8) and Lemma 3.7, it follows that $C_{min} \subseteq \text{Circ}_D(\alpha^*_{\boldsymbol{p}})$. Using Corollary 4.2, we then obtain

$$\begin{aligned} \mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] &\geq \mathbb{P}[\boldsymbol{Q}\mathcal{L}_d \cap \text{Circ}_D(\alpha^*_{\boldsymbol{p}}) \neq \{\boldsymbol{0}\}] \\ &\geq \mathbb{P}[\boldsymbol{Q}\mathcal{L}_d \cap C_{min} \neq \{\boldsymbol{0}\}] \\ &\geq 2v_{D-d+1}(C_{min}), \end{aligned} \tag{4.9}$$

where the last inequality follows from the same line of argument as in (4.6). Using (3.8) and (3.11), it is easy to verify that $2v_{D-d+1}(C_{min}) = \tau(r_{min}, d, D)$. We have shown (4.7) for $\boldsymbol{p} \in \mathcal{X} \setminus G_\epsilon$.

For $\boldsymbol{p} \in G_\epsilon$, (4.7) holds trivially, since if $\boldsymbol{p} \in G_\epsilon$, (RP$\mathcal{X}$) is ϵ-successful with probability 1. As a sanity check, $1 \geq 2v(C_{min}) = \tau(r_{min}, d, D)$ where the inequality is implied by the Gauss-Bonnet formula (3.4). □

Unfortunately, the formula defining $\tau(r, d, D)$ is not easy to interpret. To better understand the dependence of the lower bounds (4.4) and (4.7) on the parameters of the problem, we now analyse the behaviour of $\tau(r, d, D)$ in the asymptotic regime.

4.1 Asymptotic expansions

We establish the asymptotic behaviour of $\tau(r, d, D)$ for large D. The other parameters are kept fixed except for r which we allow to decrease with D. Note indeed that $r_{\boldsymbol{p}}$ in Theorem 4.3 is inversely proportional to $\|\boldsymbol{x}^* - \boldsymbol{p}\|$, which typically increases with D. Before we begin, we first need to establish the following lemma.

Lemma 4.5. *Let $0 < \alpha < \pi/2$ be either a fixed angle or a function of D that tends to 0 as $D \to \infty$. Then, as $D \to \infty$,*

$$\int_0^\alpha \sin^D(x)dx = \frac{1}{D}\frac{\sin^{D+1}(\alpha)}{\cos(\alpha)} + O\left(\frac{\sin^{D+1}(\alpha)}{D^2}\right). \tag{4.10}$$

Proof. We write

$$\int_0^\alpha \sin^D(x)dx = \int_0^\alpha \frac{\sin(x)}{D\cos(x)} \cdot (D\cos(x)\sin^{D-1}(x))dx. \tag{4.11}$$

Integration by parts with $u = \sin(x)/(D\cos(x))$ and $dv = D\cos(x)\sin^{D-1}(x)dx$ yields

$$\int_0^\alpha \sin^D(x)dx = \frac{\sin^{D+1}(\alpha)}{D\cos(\alpha)} - \frac{1}{D}\int_0^\alpha \frac{\sin^D(x)}{\cos^2(x)}dx. \tag{4.12}$$

Let I denote $\int_0^\alpha \frac{\sin^D(x)}{\cos^2(x)}dx$. It remains to show that $I = O(\sin^{D+1}(\alpha)/D)$. We express I as

$$\int_0^\alpha \frac{\sin(x)}{D\cos^3(x)} \cdot (D\cos(x)\sin^{D-1}(x))dx. \tag{4.13}$$

We integrate I by parts with $u = \sin(x)/(D\cos^3(x))$ and $dv = D\cos(x)\sin^{D-1}(x)dx$ to obtain

$$I = \frac{1}{D}\frac{\sin^{D+1}(\alpha)}{\cos^3(\alpha)} - \frac{1}{D}\int_0^\alpha \frac{1+2\sin^2(x)}{\cos^4(x)}\sin^D(x)dx \tag{4.14}$$

Since the latter integral is positive, we have

$$I \le \frac{1}{\cos^3(\alpha)} \cdot \frac{\sin^{D+1}(\alpha)}{D}. \tag{4.15}$$

Since I is positive for any $0 < \alpha < \pi/2$, (4.15) implies that $I = O(\sin^{D+1}(\alpha)/D)$. □

We establish the asymptotic behaviours of $\tau(r_{\boldsymbol{p}}, d, D)$ and $\tau(r_{min}, d, D)$ by analysing the asymptotics of $\tau(r, d, D)$ defined in (4.5) and later substituting $r_{\boldsymbol{p}}$ and r_{min} for r in $\tau(r, d, D)$.

Theorem 4.6. *Let $\tau(r, d, D)$ be defined in* (4.5)*. Let d be fixed and let r be either fixed or tend to zero as $D \to \infty$. Then,*

$$\tau(r, d, D) = \Theta\left(D^{\frac{d-2}{2}} r^{D-d}\right) \text{ as } D \to \infty, \tag{4.16}$$

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

Proof. We prove (4.16) for $d = 1$ and $1 < d < D$ separately.

First, assume that $d > 1$. By definition of $\tau(r, d, D)$, we have

$$\tau(r, d, D) = \binom{\frac{D-2}{2}}{\frac{D-d}{2}} r^{D-d}(1-r^2)^{\frac{d-2}{2}}. \tag{4.17}$$

Let us first determine the asymptotic behaviour of the binomial coefficient. Using the fact that $\Gamma(z+a)/\Gamma(z+b) = \Theta(z^{a-b})$ for large z (see, e.g., [61]), we obtain

$$\binom{\frac{D-2}{2}}{\frac{D-d}{2}} = \frac{\Gamma(\frac{D}{2})}{\Gamma(\frac{D-d+2}{2})\Gamma(\frac{d}{2})} = \frac{\Gamma(\frac{D-d+2}{2} + \frac{d-2}{2})}{\Gamma(\frac{D-d+2}{2})\Gamma(\frac{d}{2})} = \Theta\left(\left(\frac{D-d+2}{2}\right)^{\frac{d-2}{2}}\right) = \Theta\left(D^{\frac{d-2}{2}}\right). \tag{4.18}$$

To obtain[5] (4.16), we substitute (4.18) into (4.17). Note that $(1-r^2)^{\frac{d-2}{2}}$ is bounded above and bounded away from zero by constants independent of D; thus, it can be absorbed into the constants of Θ.

Let us now prove (4.16) for $d=1$. We have

$$\tau(r,d,D) = (D-1)\cdot\binom{\frac{D-2}{2}}{\frac{D-1}{2}}\int_0^{\arcsin(r)}\sin^{D-2}(x)dx, \tag{4.19}$$

where, by (4.18),

$$\binom{\frac{D-2}{2}}{\frac{D-1}{2}} = \Theta\left(D^{-\frac{1}{2}}\right) \tag{4.20}$$

and, by Lemma 4.5,

$$\int_0^{\arcsin(r)}\sin^{D-2}(x)dx = \Theta\left(\frac{1}{D-1}\frac{r^{D-1}}{\sqrt{1-r^2}}\right). \tag{4.21}$$

By substituting (4.20) and (4.21) into (4.19), we obtain (4.16) for $d=1$. For similar reasons as stated above, we can relegate the term $1/\sqrt{1-r^2}$ in (4.21) into the constants of Θ. □

Now, to obtain the asymptotics for $\tau(r_{\boldsymbol{p}},d,D)$ and $\tau(r_{\min},d,D)$, we simply apply Theorem 4.6 for $r=r_{\boldsymbol{p}}=\epsilon/(L\|\boldsymbol{x}^*-\boldsymbol{p}\|)$ and $r=r_{\min}=\epsilon/(LR_{\max})$, respectively.

Corollary 4.7. *Asymptotically for $D\to\infty$, keeping d, ϵ and L fixed and letting $\|\boldsymbol{x}^*-\boldsymbol{p}\|$ be either fixed or tend to infinity as $D\to\infty$, the lower bounds* (4.4) *and* (4.7) *satisfy*

$$\tau(r_{\boldsymbol{p}},d,D) = \Theta\left(D^{\frac{d-2}{2}}\left(\frac{\epsilon}{L\|\boldsymbol{x}^*-\boldsymbol{p}\|}\right)^{D-d}\right) \text{ as } D\to\infty, \tag{4.22}$$

with $r_{\boldsymbol{p}}=\epsilon/(L\|\boldsymbol{x}^*-\boldsymbol{p}\|)$ and where the constants in $\Theta(\cdot)$ are independent of D. Similarly,

$$\tau(r_{\min},d,D) = \Theta\left(D^{\frac{d-2}{2}}\left(\frac{\epsilon}{LR_{\max}}\right)^{D-d}\right) \text{ as } D\to\infty, \tag{4.23}$$

with $r_{\min}=\epsilon/(LR_{\max})$.

Proof. Note that $r_{\boldsymbol{p}}=\epsilon/(L\|\boldsymbol{x}^*-\boldsymbol{p}\|)$ is either fixed or tends to zero as $D\to\infty$. Then, the result follows from Theorem 4.6. □

Corollary 4.7 shows that for any $\boldsymbol{p}$ not in G_ϵ, the lower bounds in Theorem 4.3 and Theorem 4.4 decrease exponentially with D, which is as expected since problem (P) is generally NP-hard. Note that this decrease is slower for larger values of d or $\boldsymbol{p}$ closer to $\boldsymbol{x}^*$, which is reassuring.

4.2 Comparing (RP$\mathcal{X}$) to simple random search

Using the above lower bounds on the probability of ϵ-success of the reduced problem (RP$\mathcal{X}$), we now compare (RP$\mathcal{X}$) to a simple random search method to understand the relative performance of (RP$\mathcal{X}$) and when it is beneficial to use it for general functions. As a baseline for comparison, we use Uniform Sampling (US) and we restrict ourselves, in this section, to the specific case $\mathcal{X}=[-1,1]^D$ (as this will allow us to estimate the probability of success of US). We start off

[5]Here, we have also used the fact that if functions $f(x)$, $f'(x)$, $g(x)$ and $g'(x)$ satisfy $f(x)=\Theta(g(x))$ and $f'(x)=\Theta(g'(x))$ (as $x\to\infty$), then $f(x)f'(x)=\Theta(g(x)g'(x))$.

with the derivation of a lower bound for the probability of ϵ-success of US and the computation of its asymptotics.

Note that if a uniformly sampled point falls inside $B_{\epsilon/L}(\boldsymbol{x}^*)$ then US is ϵ-successful. This implies that

$$\mathbb{P}[\text{US is } \epsilon\text{-successful}] \geq \frac{\text{Vol}(B_{\epsilon/L}(\boldsymbol{x}^*))}{\text{Vol}(\mathcal{X})} = \frac{\pi^{D/2}}{2^D \Gamma(\frac{D}{2}+1)} \left(\frac{\epsilon}{L}\right)^D := \tau_{us}, \tag{4.24}$$

where we have used the fact that $\text{Vol}(B_{\epsilon/L}(\boldsymbol{x}^*)) = \frac{\pi^{D/2}}{\Gamma(\frac{D}{2}+1)} \left(\frac{\epsilon}{L}\right)^D$ (see [51, Equation 5.19.4]) and that $\text{Vol}(\mathcal{X}) = 2^D$.

Using Stirling's approximation, it is straightforward to establish the asymptotic behaviour of the lower bound τ_{us}.

Lemma 4.8. *Let τ_{us} be defined in* (4.24) *and let ϵ and L be fixed. Then,*

$$\tau_{us} = \Theta\left(D^{-\frac{D}{2}-\frac{1}{2}} \left(\frac{\pi e}{2}\right)^{\frac{D}{2}} \left(\frac{\epsilon}{L}\right)^D\right) \text{ as } D \to \infty. \tag{4.25}$$

Proof. By Stirling's approximation (see [51, Equation 5.11.7]),

$$\Gamma\left(\frac{D}{2}+1\right) = \Theta\left(e^{-\frac{D}{2}} \left(\frac{D}{2}\right)^{\frac{D+1}{2}}\right) \text{ as } D \to \infty. \tag{4.26}$$

By substituting (4.26) into (4.24), we obtain the desired result. □

Let us now compare the lower bound τ_{us} of US to the lower bound $\tau(r_{\boldsymbol{p}}, d, D)$ for (RP$\mathcal{X}$). It is clear from the analysis of $\tau(r_{\boldsymbol{p}}, d, D)$ in Section 4.1 that the probability of ϵ-success of (RP$\mathcal{X}$) is higher if $\boldsymbol{p}$ is closer to the set of global minimizers. In the next theorem, we determine a threshold distance Δ_0 between $\boldsymbol{p}$ and a global minimizer $\boldsymbol{x}^*$ such that $\tau(r_{\boldsymbol{p}}, d, D)$ and τ_{us} are approximately equal to each other. This would tell us how close $\boldsymbol{p}$ should be to $\boldsymbol{x}^*$ for (RP$\mathcal{X}$) to have a larger lower bound for the probability of success than that of US. The analysis is done in the asymptotic regime.

Theorem 4.9. *Suppose that Assumption LipC holds, and that $\mathcal{X} = [-1,1]^D$. Let $\boldsymbol{x}^*$ be a global minimizer of* (P) *satisfying Assumption FeasBall. Let $\tau(r_{\boldsymbol{p}}, d, D)$ and τ_{us} be defined in Theorem 4.3 and* (4.24)*, respectively. Let ϵ, L, d be fixed and let $\Delta_0 = \sqrt{\frac{2D}{\pi e}}$. Then,*

a) If $\lim_{D\to\infty} \frac{\Delta_0}{\|\boldsymbol{x}^* - \boldsymbol{p}\|} = \psi > 1$, then $\tau(r_{\boldsymbol{p}}, d, D)/\tau_{us} \to \infty$ as $D \to \infty$.

b) If $\lim_{D\to\infty} \frac{\Delta_0}{\|\boldsymbol{x}^* - \boldsymbol{p}\|} = \psi < 1$, then $\tau(r_{\boldsymbol{p}}, d, D)/\tau_{us} \to 0$ as $D \to \infty$.

Proof. From (4.23) and (4.25), we have

$$\begin{aligned}\frac{\tau(r_{\boldsymbol{p}}, d, D)}{\tau_{us}} = \frac{\Theta\left(D^{\frac{d-2}{2}} \left(\frac{\epsilon}{L\|\boldsymbol{x}^*-\boldsymbol{p}\|}\right)^{D-d}\right)}{\Theta\left(\frac{1}{\sqrt{D}} \left(\frac{\pi e}{2D}\right)^{\frac{D}{2}} \left(\frac{\epsilon}{L}\right)^D\right)} &\overset{6}{=} \Theta\left(\left(\frac{\epsilon}{L}\right)^{-d} \left(\frac{2}{\pi e}\right)^{D/2} D^{\frac{D+d-1}{2}} \|\boldsymbol{x}^* - \boldsymbol{p}\|^{d-D}\right) \\ &= \Theta\Bigg(\Bigg(\underbrace{\left[\frac{\sqrt{2D/\pi e}}{\|\boldsymbol{x}^* - \boldsymbol{p}\|}\right]}_{= \Delta_0/\|\boldsymbol{x}^* - \boldsymbol{p}\|} \cdot D^{\frac{2d-1}{2(D-d)}}\Bigg)^{D-d}\Bigg),\end{aligned} \tag{4.27}$$

[6] Here, we use the fact that if functions $f(x)$, $f'(x)$, $g(x)$ and $g'(x)$ satisfy $f(x) = \Theta(g(x))$ and $f'(x) = \Theta(g'(x))$ (as $x \to \infty$), then $f(x)/f'(x) = \Theta(g(x)/g'(x))$.

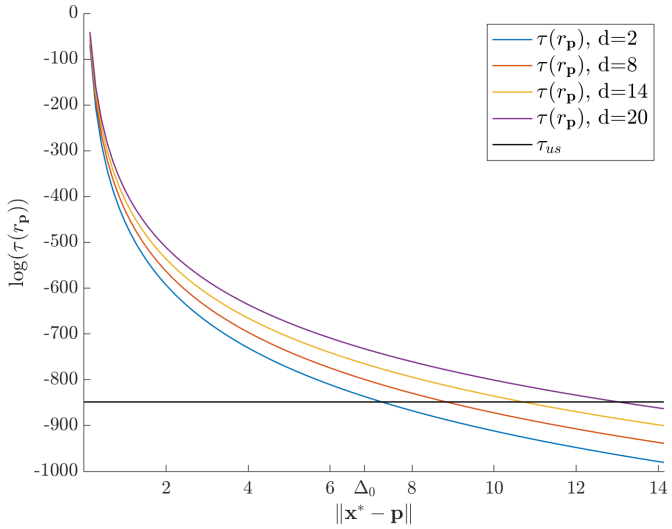

Figure 3: A plot of $\tau(r_{\boldsymbol{p}})$ versus $\|\boldsymbol{x}^* - \boldsymbol{p}\|$ for different values of the subspace embedding dimension d. The lower bound τ_{us} of US does not depend on $\|\boldsymbol{x}^* - \boldsymbol{p}\|$ and, thus, it is displayed as a straight horizontal line.

Note that in the second line there is a term $\left(\frac{\epsilon}{L}\right)^{-d}\left(\frac{2}{\pi e}\right)^{d/2}$ missing inside Θ, which we removed as it is independent of D. Now, by definition of Θ, (4.27) implies that there exist positive constants M_1 and M_2 such that

$$M_1\left(\frac{\Delta_0}{\|\boldsymbol{x}^* - \boldsymbol{p}\|}D^{\frac{2d-1}{2(D-d)}}\right)^{D-d} \leq \frac{\tau(r_{\boldsymbol{p}}, d, D)}{\tau_{us}} \leq M_2\left(\frac{\Delta_0}{\|\boldsymbol{x}^* - \boldsymbol{p}\|}D^{\frac{2d-1}{2(D-d)}}\right)^{D-d} \tag{4.28}$$

as $D \to \infty$. Note that $D^{\frac{2d-1}{2(D-d)}} \to 1$ as $D \to \infty$. Hence, if $\Delta_0/\|\boldsymbol{x}^* - \boldsymbol{p}\| \to \psi > 1$ then both lower and upper bounds in (4.28) tend to infinity implying that $\tau(r_{\boldsymbol{p}}, d, D)/\tau_{us} \to \infty$. On the other hand, if $\Delta_0/\|\boldsymbol{x}^* - \boldsymbol{p}\| \to \psi < 1$ then both lower and upper bounds in (4.28) tend to zero implying that $\tau(r_{\boldsymbol{p}}, d, D)/\tau_{us} \to 0$. □

Theorem 4.9 tells us that the distance between $\boldsymbol{p}$ and $\boldsymbol{x}^*$ (in the asymptotic setting) must be no greater than $\Delta_0 \approx 0.48\sqrt{D}$ for $\tau(r_{\boldsymbol{p}}, d, D)$ to be larger than τ_{us} in the case $\mathcal{X} = [-1, 1]^D$. Note that, since the distance between the origin and a corner of $\mathcal{X}$ is equal to $\sqrt{D}$ $(> 0.48\sqrt{D})$, there is no point $\boldsymbol{p}$ such that the ball of radius Δ_0 centred at $\boldsymbol{p}$ covers all points in $\mathcal{X}$. In other words, in the specific case $\mathcal{X} = [-1, 1]^D$, for any $\boldsymbol{p}$ in $\mathcal{X}$, there always exists $\boldsymbol{x}^*$ for which $\tau(r_{\boldsymbol{p}}, d, D)$ is smaller than τ_{us}; on the other hand, if $\boldsymbol{p} = \boldsymbol{0}$ and $\boldsymbol{x}^*$ is close to the origin then $\tau(r_{\boldsymbol{p}}, d, D) > \tau_{us}$. Note also that Δ_0 has no dependence on the embedding subspace dimension d. This is due to the asymptotic nature of the analysis: in (4.28), we see that both inequalities depend on d, but the dependence diminishes as $D \to \infty$ since d is kept fixed. Although the asymptotic analysis shows no significant dependence on the subspace dimension, numerical experiments show that the value of d has a notable effect on success of (RP$\mathcal{X}$). In Figure 3, we plot $\tau(r_{\boldsymbol{p}}, d, D)$ as a function of $\|\boldsymbol{x}^* - \boldsymbol{p}\|$ for different values of d with D fixed at 200. The lower bound τ_{us} of US is represented by a black horizontal line. We see that, for larger d, $\tau(r_{\boldsymbol{p}}, d, D)$ decreases at a slower rate and has greater threshold distance before becoming smaller than τ_{us}.

Remark 4.10. An important distinction must be made between the implications of the ϵ-success of (RP$\mathcal{X}$) and the ϵ-success of US in solving the original problem (P). Note that the ϵ-success of US means that US has sampled a point that lies in G_ϵ, which in turn implies that US has

successfully (approximately) solved (P). This is not the case for (RP$\mathcal{X}$). Recall that ϵ-success of (RP$\mathcal{X}$) by definition means that there is an approximate solution $\boldsymbol{x}^*$ to (P) that lies in the embedded d-dimensional subspace. One needs to perform an additional global search over the subspace to locate $\boldsymbol{x}^*$. Therefore, for an entirely fair comparison between the two approaches, this additional computational complexity should be taken into account.

5 X-REGO: an algorithmic framework for global optimization using random embeddings

This section presents the proposed algorithmic framework for global optimization using random embeddings, named X-REGO by analogy with [18] (see the Introduction for distinctions between these variants). X-REGO is a generic algorithmic framework that replaces the high-dimensional original problem (P) by a sequence of low-dimensional random problems of the form (RP$\mathcal{X}$); these reduced random problems can then be solved using any global — and in practice, even a local — optimization solver.

Note that the kth embedding in X-REGO is determined by a realization $\tilde{\boldsymbol{A}}^k = \boldsymbol{A}^k(\boldsymbol{\omega}^k)$ of the random Gaussian matrix $\boldsymbol{A}^k \in \mathbb{R}^{D\times d^k}$, for some (deterministic) $d^k \in \{1,\dots,D-1\}$. For generality of our analysis, we also assume that the parameter $\boldsymbol{p}$ in (RP$\mathcal{X}$) is a random variable. The kth embedding is drawn at the point $\tilde{\boldsymbol{p}}^{k-1} = \boldsymbol{p}^{k-1}(\boldsymbol{\omega}^{k-1})$, a realization of the random variable $\boldsymbol{p}^{k-1}$, assumed to have support included in $\mathcal{X}$. Note that this definition includes deterministic choices for $\boldsymbol{p}^{k-1}$, by writing it as a random variable with support equal to a singleton (deterministic and stochastic selection rules for the $\boldsymbol{p}$ are given below).

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

1: Initialize $d^1 \in \{1,2,\dots,D-1\}$ and $\tilde{\boldsymbol{p}}^0 \in \mathcal{X}$.
2: **for** $k \geq 1$ until termination **do**
3: Draw $\tilde{\boldsymbol{A}}^k$, a realization of the $D \times d^k$ Gaussian matrix $\boldsymbol{A}^k$.
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

$$\tilde{\boldsymbol{x}}^k := \tilde{\boldsymbol{A}}^k\tilde{\boldsymbol{y}}^k + \tilde{\boldsymbol{p}}^{k-1}. \qquad (5.1)$$

6: Choose (deterministically or randomly) $\tilde{\boldsymbol{p}}^k \in \mathcal{X}$.
7: Choose $d^{k+1} \in \{1,2,\dots,D-1\}$.
8: **end for**

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

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

As such, X-REGO can be seen as a stochastic process: additionally 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. To calculate

$\tilde{\boldsymbol{y}}^k$, ($\widetilde{\mathrm{RP}\mathcal{X}^k}$) may be solved to some required accuracy using a deterministic global optimization algorithm that is allowed to fail with a certain probability; or employing a stochastic algorithm, so that $\tilde{\boldsymbol{y}}^k$ is only guaranteed to be an approximate global minimizer of ($\widetilde{\mathrm{RP}\mathcal{X}^k}$) (at least) with a certain probability. This allows us to account for solvers having some stochastic component (multistart methods, genetic algorithms, ...), or deterministic solvers that may fail in some cases due, e.g., to a computational budget shortage.

Note also that the choice of the random variable $\boldsymbol{p}^k$ and of the subspace dimension d^k provide some flexibility in the algorithm. For $\boldsymbol{p}^k$, possibilities include:

- $\boldsymbol{p}^k = \boldsymbol{p}$: all the random embeddings explored are drawn at the same point (in case $\boldsymbol{p}$ is a fixed vector in $\mathcal{X}$), or according to the same distribution (if $\boldsymbol{p}$ is a random variable),
- The sequence $\boldsymbol{p}^0, \boldsymbol{p}^1, \ldots$ can be constructed dynamically during the optimization, e.g., based on the information gathered so far on the objective. For example, one may choose $\boldsymbol{p}^k = \boldsymbol{x}^k_{opt}$, where $\boldsymbol{x}^k_{opt}$ is the best point found up to the kth embedding:

$$\boldsymbol{x}^k_{opt} := \arg\min\{f(\boldsymbol{x}^1), f(\boldsymbol{x}^2), \ldots, f(\boldsymbol{x}^k)\}. \tag{5.2}$$

Note that ($\widetilde{\mathrm{RP}\mathcal{X}^k}$) is always feasible for all choices of $\boldsymbol{p}^k$ ($\boldsymbol{y} = 0$ is feasible since $\tilde{\boldsymbol{p}}^k \in \mathcal{X}$). However, it may happen that this is the only feasible point of ($\widetilde{\mathrm{RP}\mathcal{X}^k}$); to avoid this situation we may assume that $\boldsymbol{p}^k$ is in the interior of $\mathcal{X}$. This latter assumption is not needed for our convergence results to hold, but it is a desirable assumption from a numerical point of view.

Regarding the subspace dimension d^k, one can be for example choose a fixed value based on the computational budget available for the reduced problem, or d^k can be progressively increased, using a warm start in each embedding. We refer the reader to Section 8 for a numerical comparison of some of those strategies.

The termination in Line 2 could be set to a given maximum number of embeddings, or could check that no significant progress in decreasing the objective function has been achieved over the last few embeddings, compared to the value $f(\tilde{\boldsymbol{x}}^k_{opt})$. For generality, we leave it unspecified here.

6 Global convergence of X-REGO to a set of global ϵ-minimizers

The convergence results presented in this paper extend the ones given in [18], in which X-REGO (with fixed subspace dimension $d^k = d \geq d_e$ for all k) is proven to converge for functions with low-effective dimension d_e. Section 6.1 is devoted to a generic convergence analysis of X-REGO, under generic assumptions on the probability of ϵ-success of ($\mathrm{RP}\mathcal{X}^k$) and on the probability of success of the solver to find an approximate minimizer of its realisation ($\widetilde{\mathrm{RP}\mathcal{X}^k}$), while Section 6.2 presents the application of these results to arbitrary Lipschitz-continuous objectives, building on the results presented in the previous sections to show the validity of the ϵ-success assumption.

6.1 A general convergence framework for X-REGO

This section recalls results in [18] that are needed for our main convergence results in the next section. We show that $\boldsymbol{x}^k_{opt}$ defined in (5.2) converges to the set of ϵ-minimizers G_ϵ almost surely as $k \to \infty$ (see Theorem 6.3). Intuitively, our proof relies on the fact that any vector $\tilde{\boldsymbol{x}}^k$ defined in (5.1) belongs to G_ϵ if the following two conditions hold simultaneously:

(a) the reduced problem ($\mathrm{RP}\mathcal{X}^k$) is $(\epsilon - \lambda)$-successful in the sense of Definition 1.1, namely,

$$f^k_{min} \leq f^* + \epsilon - \lambda; \tag{6.1}$$

(b) the reduced problem $(\widetilde{\mathrm{RP}\mathcal{X}^k})$ is solved (by a deterministic/stochastic algorithm) to an accuracy $\lambda \in (0, \epsilon)$ in the objective function value, namely,

$$f(\boldsymbol{A}^k \boldsymbol{y}^k + \boldsymbol{p}^{k-1}) \leq f^k_{min} + \lambda \tag{6.2}$$

holds (at least) with a certain probability.

Note that in order to prove convergence of X-REGO to (global) ϵ-minimizers, the value of ϵ in the success probability of the reduced problem $(\mathrm{RP}\mathcal{X})$ needs to be replaced by $(\epsilon - \lambda)$. This change is motivated by the fact that we allow inexact solutions (up to accuracy λ) of the reduced problem $(\widetilde{\mathrm{RP}\mathcal{X}^k})$. We also emphasize that, according to the discussion in Section 5, and for the sake of generality, the parameter $\boldsymbol{p}^k$ in $(\mathrm{RP}\mathcal{X}^k)$ is now a random variable (in contrast with Section 4 where it was assumed to be deterministic).

Let us introduce two additional random variables that capture the conditions in (a) and (b) above,

$$R^k = \mathbb{1}\{(\mathrm{RP}\mathcal{X}^k) \text{ is } (\epsilon - \lambda)\text{-successful in the sense of (6.1)}\}, \tag{6.3}$$

$$S^k = \mathbb{1}\{(\mathrm{RP}\mathcal{X}^k) \text{ is solved to accuracy } \lambda \text{ in the sense of (6.2)}\}, \tag{6.4}$$

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)[7], 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[8], and R^k is also $\mathcal{F}^{k-1/2}$-measurable; thus they are well-defined random variables.

Remark 6.1. 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 $\mathcal{F}^0 \subseteq \mathcal{F}^{1/2} \subseteq \mathcal{F}^1 \subseteq \cdots \subseteq \mathcal{F}^{k-1} \subseteq \mathcal{F}^{k-1/2}$.

The following assumption says that the reduced problem $(\mathrm{RP}\mathcal{X}^k)$ needs to be solved to required accuracy with some positive probability. Note that this is a rather weak assumption, that is satisfied by any reasonable solver.

Assumption Success-Solv. For all $k \geq 1$, there exists $\rho^k \in [\rho_{lb}, 1]$, with $\rho_{lb} > 0$ such that[9]

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

i.e., with (conditional) probability at least $\rho^k \geq \rho_{lb}$, the solution $\boldsymbol{y}^k$ of $(\mathrm{RP}\mathcal{X}^k)$ satisfies (6.2).[10]

[7] A similar setup regarding random iterates of probabilistic models can be found in [5, 17] in the context of local optimization.

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

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

[10] In general, ρ^k will depend on the dimension d^k of the kth random embedding.

Remark 6.2. If a deterministic (global optimization) algorithm is used to solve $(\widetilde{\mathrm{RP}\mathcal{X}^k})$, then S^k is always $\mathcal{F}_k^{k-1/2}$-measurable and Assumption Success-Solv is equivalent to $S^k \geq \rho^k > 0$. Since S^k is an indicator function, this further implies that $S^k \equiv 1$.

The next assumption says that the drawn subspaces are $(\epsilon - \lambda)$-successful with a positive probability.

Assumption Succes-Emb. For all $k \geq 1$, there exists $\tau^k \in [\tau_{lb}, 1]$, with $\tau_{lb} > 0$ such that

$$\mathbb{P}[R^k = 1 | \mathcal{F}^{k-1}] = \mathbb{E}[R^k | \mathcal{F}^{k-1}] \geq \tau^k, \tag{6.5}$$

i.e., with (conditional) probability at least $\tau^k \geq \tau_{lb} > 0$, $(\mathrm{RP}\mathcal{X}^k)$ is $(\epsilon - \lambda)$-successful.

Note that Assumption Success-Solv and Assumption Succes-Emb have been slightly modified compared to [18]: here, the dimension of the reduced problem is varying, so in general the probabilities of success of the solver and embedding depend on k as well. Under Assumption Success-Solv and Assumption Succes-Emb, the following result shows the convergence of X-REGO to the set of ϵ-minimizers.

Theorem 6.3 (Global convergence)**.** *Suppose Assumption Success-Solv and Succes-Emb hold. Then,*

$$\lim_{k\to\infty} \mathbb{P}[\boldsymbol{x}_{opt}^k \in G_\epsilon] = \lim_{k\to\infty} \mathbb{P}[f(\boldsymbol{x}_{opt}^k) \leq f^* + \epsilon] = 1 \tag{6.6}$$

where $\boldsymbol{x}_{opt}^k$ and G_ϵ are defined in (5.2) *and* (2.1)*, respectively. Furthermore, for any $\xi \in (0, 1)$,*

$$\mathbb{P}[\boldsymbol{x}_{opt}^k \in G_\epsilon] = \mathbb{P}[f(\boldsymbol{x}_{opt}^k) \leq f^* + \epsilon] \geq \xi \text{ for all } k \geq K_\xi, \tag{6.7}$$

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

Proof. The proof is a straightforward extension of the one given in [18], and for completeness, we include it in Appendix B.1. □

Remark 6.4. If the original problem (P) is convex (and known a priori to be so), then clearly, a local (deterministic or stochastic) optimization algorithm may be used to solve $(\widetilde{\mathrm{RP}\mathcal{X}^k})$ and achieve (6.2). Apart from this important speed-up and simplification, it seems difficult at present to see how else problem convexity could be exploited in order to improve the success bounds and convergence of X-REGO.

6.2 Global convergence of X-REGO for general objectives

The previous section provides a convergence result, with associate convergence rate, that depends on some parameters ρ_{lb} and τ_{lb} defined in Assumption Success-Solv and Succes-Emb. The former intrinsically depends on the solver used to solve the reduced subproblems, and will not be discussed further here. However, the latter parameter τ_{lb} can be estimated for general Lipschitz-continuous objectives using the results derived in Section 4.

Corollary 6.5. *Suppose that Assumption LipC holds, that there exists a global minimizer $\boldsymbol{x}^*$ of* (P) *that satisfies Assumption FeasBall (replacing ϵ by $\epsilon - \lambda$ in Assumption FeasBall, i.e., slightly relaxing the assumption), and that $\tilde{\boldsymbol{p}}^k$ satisfies $\|\tilde{\boldsymbol{p}}^k - \boldsymbol{x}^*\| \leq R_{\max}$ for all k and for some suitably chosen $R_{\max}$. Suppose also that $d^k \geq d_{lb}$ for some $d_{lb} > 0$. Then, Assumption Succes-Emb holds with*

$$\tau_{lb} = \tau(r_{\min}, d_{lb}, D),$$

with $r_{\min} = (\epsilon - \lambda)/(LR_{\max})$.

Proof. Let us first recall that for all k, there holds by Corollary 4.2:

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}^k) \text{ is } (\epsilon-\lambda)\text{-successful}] \geq \mathbb{P}[\boldsymbol{Q}\mathcal{L}_{d^k} \cap \mathrm{Circ}_D(\alpha^*_{\tilde{\boldsymbol{p}}^{k-1}}) \neq \{\mathbf{0}\}],$$

where $\boldsymbol{Q}$ is a $D \times D$ random orthogonal matrix drawn uniformly from the set of all $D \times D$ real orthogonal matrices, $\mathcal{L}_{d^k}$ a d^k-dimensional linear subspace, and $\alpha^*_{\tilde{\boldsymbol{p}}^{k-1}} := \arcsin((\epsilon-\lambda)/\|\boldsymbol{x}^* - \tilde{\boldsymbol{p}}^{k-1}\|)$. Let $\alpha^*_{\min} := \arcsin((\epsilon-\lambda)/(LR_{\max}))$, and note that $\alpha^*_{\min} \leq \alpha^*_{\tilde{\boldsymbol{p}}^{k-1}}$ for all k. By Lemma 3.7, for any $\alpha^*_{\min} \leq \alpha \leq \pi/2$, there holds $\mathrm{Circ}_D(\alpha^*_{\min}) \subseteq \mathrm{Circ}_D(\alpha)$ so that

$$\mathbb{P}[\boldsymbol{Q}\mathcal{L}_{d^k} \cap \mathrm{Circ}_D(\alpha^*_{\tilde{\boldsymbol{p}}^{k-1}}) \neq \{\mathbf{0}\}] \geq \mathbb{P}[\boldsymbol{Q}\mathcal{L}_{d^k} \cap \mathrm{Circ}_D(\alpha^*_{\min}) \neq \{\mathbf{0}\}]$$

for all k. By the Crofton formula, there holds

$$\mathbb{P}[\boldsymbol{Q}\mathcal{L}_{d^k} \cap \mathrm{Circ}_D(\alpha^*_{\min}) \neq \{\mathbf{0}\}] = 2h_{D-d^k+1}.$$

By [3, Prop. 5.9], $h_k \geq h_{k+1}$ for all $k = 0, \ldots, D-1$. We deduce that

$$\mathbb{P}[\boldsymbol{Q}\mathcal{L}_{d^k} \cap \mathrm{Circ}_D(\alpha^*_{\min}) \neq \{\mathbf{0}\}] = 2h_{D-d^k+1} \geq 2h_{D-d_{lb}+1}.$$

Using the fact that the intrinsic volumes are all non-negative, and the definition of h_k, we get:

$$\mathbb{P}[\boldsymbol{Q}\mathcal{L}_{d^k} \cap \mathrm{Circ}_D(\alpha^*_{\min}) \neq \{\mathbf{0}\}] \geq 2v_{D-d_{lb}+1} = \tau(r_{\min}, d_{lb}, D).$$

Note finally that, in terms of conditional expectation, we can write $\mathbb{E}[R^k|\mathcal{F}^{k-1}] = 1 \cdot \mathbb{P}[R^k = 1|\mathcal{F}^{k-1}] + 0 \cdot \mathbb{P}[R^k = 0|\mathcal{F}^{k-1}] \geq \tau_{lb}$. This shows that (6.5) in Assumption Succes-Emb holds. □

We now estimate the rate of convergence of X-REGO for Lipschitz continuous functions using the estimates for τ provided in Corollary 4.7.

Theorem 6.6. *Suppose that Assumptions LipC and Success-Solv hold, that there exists a global minimizer $\boldsymbol{x}^*$ of* (P) *that satisfies Assumption FeasBall (replacing ϵ by $\epsilon-\lambda$ in Assumption FeasBall), and that $\tilde{\boldsymbol{p}}^k$ satisfies $\|\tilde{\boldsymbol{p}}^k - \boldsymbol{x}^*\| \leq R_{\max}$ for all k and for some suitably chosen $R_{\max}$. Suppose also that $d^k \geq d_{lb}$ for some $d_{lb} > 0$. Then, $\boldsymbol{x}^k_{opt}$ defined in* (5.2) *converges to the set of ϵ-minimizers G_ϵ almost surely as $k \to \infty$, and*

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

with

$$K_\xi = \frac{|\log(1-\xi)|}{\rho_{lb}} O\left(D^{\frac{2-d_{lb}}{2}} \left(\frac{LR_{\max}}{\epsilon-\lambda}\right)^{D-d_{lb}}\right) \text{ as } D \to \infty. \tag{6.8}$$

Proof. The result follows from Theorem 6.3, Corollary 6.5 and Corollary 4.7. □

6.3 Ensuring boundedness of $\tilde{\boldsymbol{p}}^k$

So far, our convergence results rely on the assumption that, for each k, $\|\tilde{\boldsymbol{p}}^k - \boldsymbol{x}^*\| \leq R_{\max}$ for some suitably chosen $R_{\max}$ and for some global minimizer $\boldsymbol{x}^*$ surrounded by a ball of radius $(\epsilon-\lambda)$ of feasible solutions, see Assumption FeasBall. We show in this section that the following strategies for choosing the random variable $\boldsymbol{p}^k$ guarantee that $\boldsymbol{x}^k_{opt}$ defined in (5.2) converges to the set of ϵ-minimizers G_ϵ almost surely as $k \to \infty$.

1. $\boldsymbol{p}^k$ is deterministic and does not vary with k (e.g., $\boldsymbol{p}^k = \mathbf{0}$ for all k).
2. $(\boldsymbol{p}^k)_{k=1,2,\ldots}$ is a bounded sequence of deterministic values.

3. $\boldsymbol{p}^k$ is any random variable with support contained in $\mathcal{X}$, and $\mathcal{X}$ is bounded.

4. $\boldsymbol{p}^k$ is a random variable defined as $\boldsymbol{p}^k = \boldsymbol{x}^k_{opt}$, where $\boldsymbol{x}^k_{opt}$ is the best point found over the k first embeddings, see (5.2), and the objective is coercive.

Note that for the strategies 1, 2 and 3, the validity of Theorem 6.6 follows simply from the triangular inequality:

$$\|\tilde{\boldsymbol{p}}^k - \boldsymbol{x}^*\| \leq \|\tilde{\boldsymbol{p}}^k\| + \|\boldsymbol{x}^*\| < \infty,$$

and the fact that $\|\tilde{\boldsymbol{p}}^k\|$ is bounded. We prove next that $\boldsymbol{x}^k_{opt}$ defined in (5.2) converges to the set of ϵ-minimizers G_ϵ almost surely as $k \to \infty$ for strategy 4 if the objective is coercive.

Assumption 6.7 (Coerciveness, see [6])**.** *When $\mathcal{X}$ is unbounded, the (continuous) function $f : \mathcal{X} \to \mathbb{R}$ in* (P) *satisfies*

$$\lim_{\|\boldsymbol{x}\| \to \infty} f(\boldsymbol{x}) = \infty. \tag{6.9}$$

Corollary 6.8. *Let Assumption 6.7 hold, and let $\boldsymbol{x}^*$ be a global minimizer of* (P)*. Let $\boldsymbol{p}^k = \boldsymbol{x}^k_{opt}$ for $k \geq 1$, with $\boldsymbol{x}^k_{opt}$ defined in* (5.2)*, and let $\boldsymbol{p}^0 \in \mathcal{X}$ be such that $f(\tilde{\boldsymbol{p}}^0) < \infty$. Then, there exists $R_{\max}$ such that, for all k,*

$$\|\tilde{\boldsymbol{p}}^k - \boldsymbol{x}^*\| \leq R_{\max}. \tag{6.10}$$

Proof. Note that the sequence $(f(\tilde{\boldsymbol{p}}^k))_{k=0,1,2,\ldots}$ is decreasing by definition of the random variable $\boldsymbol{x}^k_{opt}$. Therefore, for all k there holds

$$f(\tilde{\boldsymbol{p}}^k) \leq f(\tilde{\boldsymbol{p}}^0) < \infty.$$

By coerciveness of f, there exists $R < \infty$ such that for any deterministic vector $\boldsymbol{y} \in \mathcal{X}$, $\|\boldsymbol{y}\| > R$ implies $f(\boldsymbol{y}) > f(\tilde{\boldsymbol{p}}^0)$. We deduce that $\|\tilde{\boldsymbol{p}}^k\| < R$ for all k, so that $\|\tilde{\boldsymbol{p}}^k - \boldsymbol{x}^*\| \leq \|\tilde{\boldsymbol{p}}^k\| + \|\boldsymbol{x}^*\| \leq R + \|\boldsymbol{x}^*\|$. The result follows by writing $R_{\max} = R + \|\boldsymbol{x}^*\|$. □

Corollary 6.9. *Suppose that Assumptions LipC, Success-Solv and 6.7 hold, that there exists a global minimizer $\boldsymbol{x}^*$ of* (P) *that satisfies Assumption FeasBall (replacing ϵ by $\epsilon - \lambda$ in Assumption FeasBall), and that $d^k \geq d_{lb}$ for some $d_{lb} > 0$. Let $\boldsymbol{p}^k = \boldsymbol{x}^k_{opt}$ for $k \geq 1$, with $\boldsymbol{x}^k_{opt}$ defined in* (5.2)*, and let $\boldsymbol{p}^0 \in \mathcal{X}$ be such that $f(\tilde{\boldsymbol{p}}^0) < \infty$. Then, $\boldsymbol{x}^k_{opt}$ converges to the set of ϵ-minimizers G_ϵ almost surely as $k \to \infty$, and there exists $R_{\max}$ such that*

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

with

$$K_\xi = \frac{|\log(1-\xi)|}{\rho_{lb}} O\left(D^{\frac{2-d_{lb}}{2}} \left(\frac{L R_{\max}}{\epsilon - \lambda} \right)^{D - d_{lb}} \right) \text{ as } D \to \infty. \tag{6.11}$$

Proof. The result follows from Theorem 6.6 and Corollary 6.8. □

7 Applying X-REGO to functions with low effective dimensionality

The recent works [15, 18] explore random embedding algorithms for functions with low effective dimension, that only vary over a subspace of dimension $d_e < D$, and address respectively the case $\mathcal{X} = \mathbb{R}^D$ and $\mathcal{X} = [-1, 1]^D$. Both papers assume that the dimension of the random subspace d in (RP$\mathcal{X}$) is the same or exceeds the effective dimension d_e, and derive bounds on the probability of (RP$\mathcal{X}$) to be ϵ-successful in that setting; these bounds are then used to prove convergence of respective random subspace methods. For the remainder of this paper, we explore the use of X-REGO for unconstrained global optimization of functions with low effective dimension, for any random subspace dimension d, thus removing the assumption $d \geq d_e$. To prove convergence of X-REGO in that setting, we rely on the results derived in Section 4.

7.1 Definitions and existing results

Definition 7.1 (Functions with low effective dimensionality, see [66])**.** A function $f : \mathbb{R}^D \to \mathbb{R}$ has effective dimension $d_e < D$ 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

$$f(\boldsymbol{x}_\top + \boldsymbol{x}_\perp) = f(\boldsymbol{x}_\top), \tag{7.1}$$

and d_e is the smallest integer satisfying (7.1).

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

Assumption LowED. The function $f : \mathbb{R}^D \to \mathbb{R}$ has effective dimensionality d_e with effective subspace[11] $\mathcal{T}$ and constant subspace $\mathcal{T}^\perp$ spanned by the columns of the orthonormal matrices $\boldsymbol{U} \in \mathbb{R}^{D\times d_e}$ and $\boldsymbol{V} \in \mathbb{R}^{D\times(D-d_e)}$, respectively. We write $\boldsymbol{x}_\top = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{x}$ and $\boldsymbol{x}_\perp = \boldsymbol{V}\boldsymbol{V}^T\boldsymbol{x}$, the unique Euclidean projections of any vector $\boldsymbol{x} \in \mathbb{R}^D$ onto $\mathcal{T}$ and $\mathcal{T}^\perp$, respectively.

As discussed in [18], functions with low effective dimension have the nice property that their global minimizers are not isolated: to any global minimizer $\boldsymbol{x}^*$ of (P), with Euclidean projection $\boldsymbol{x}^*_\top$ on the effective subspace $\mathcal{T}$, one can associate a subspace $\mathcal{G}^*$ on which the objective reaches its minimal value. Indeed, writing

$$\mathcal{G}^* = \{\boldsymbol{x}^*_\top + \boldsymbol{V}\boldsymbol{h} : \boldsymbol{h} \in \mathbb{R}^{D-d_e}\}, \tag{7.2}$$

Assumption LowED implies that $f(\boldsymbol{x}) = f^*$ for all $\boldsymbol{x} \in \mathcal{G}^*$.

In the case $d \geq d_e$, the following result, derived in [66], says that the reduced problem (RP$\mathcal{X}$) is successful with probability one.

Theorem 7.2. *(see [66, Theorem 2], and [52, Rem. 2.22]) Let $\mathcal{X} = \mathbb{R}^D$ and let Assumption LowED hold. Let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix with $d \geq d_e$, and let $\boldsymbol{p} \in \mathbb{R}^D$. Then, with probability one, for any fixed $\boldsymbol{x} \in \mathbb{R}^D$, there exists a $\boldsymbol{y} \in \mathbb{R}^d$ such that $f(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{y} + \boldsymbol{p})$. In particular, for any global minimizer $\boldsymbol{x}^*$ of* (P)*, with probability one, there exists $\boldsymbol{y}^* \in \mathbb{R}^d$ such that $f(\boldsymbol{A}\boldsymbol{y}^* + \boldsymbol{p}) = f(\boldsymbol{x}^*) = f^*$.*

Thus, in the unconstrained case $\mathcal{X} = \mathbb{R}^D$, the solution of a single reduced problem (RP$\mathcal{X}$) with subspace dimension $d \geq d_e$ provides an exact global minimizer of the original problem (P) with probability one. In the next section, we address the case $d < d_e$.

7.2 Probability of success of the reduced problem for lower dimensional embeddings

Unfortunately, Theorem 7.2 crucially depends on the assumption $d \geq d_e$. When $d < d_e$, we quantify the probability of the random embedding to contain a (global) ϵ-minimizer. Similarly to the definition of $\mathcal{G}^*$ above, one may associate to any global minimizer $\boldsymbol{x}^*$ of (P) a connected set $\mathcal{G}^*_\epsilon$ of ϵ-minimizers. Denoting the Euclidean projection of $\boldsymbol{x}^*$ on the effective subspace by $\boldsymbol{x}^*_\top$, under Assumption LipC (Lipschitz continuity of f), $\mathcal{G}^*_\epsilon$ is the Cartesian product of a d_e-dimensional ball (contained in the effective subspace) by the constant subspace $\mathcal{T}^\perp$ (see Assumption LowED):

$$\mathcal{G}^*_\epsilon := \{\boldsymbol{x}^*_\top + \boldsymbol{U}\boldsymbol{g} + \boldsymbol{V}\boldsymbol{h} : \boldsymbol{g} \in \mathbb{R}^{d_e}, \|\boldsymbol{g}\| \leq \epsilon/L, \boldsymbol{h} \in \mathbb{R}^{D-d_e}\}, \tag{7.3}$$

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

where L is the Lipschitz constant of f. Indeed, let $\boldsymbol{x} := \boldsymbol{x}^*_\top + \boldsymbol{U}\boldsymbol{g} + \boldsymbol{V}\boldsymbol{h} \in \mathcal{G}^*_\epsilon$, for some $\boldsymbol{g} \in \mathbb{R}^{d_e}$ satisfying $\|\boldsymbol{g}\| \leq \epsilon/L$ and for some $\boldsymbol{h} \in \mathbb{R}^{D-d_e}$. Then, $f(\boldsymbol{x}) = f(\boldsymbol{x}^*_\top + \boldsymbol{U}\boldsymbol{g})$ by Assumption LowED, since $\boldsymbol{V}\boldsymbol{h} \in \mathcal{T}^\perp$. By Lipschitz continuity of f, we get:

$$f(\boldsymbol{x}) = f(\boldsymbol{x}^*_\top + \boldsymbol{U}\boldsymbol{g}) \leq f(\boldsymbol{x}^*_\top) + L\|\boldsymbol{U}\boldsymbol{g}\| \leq f^* + \epsilon. \tag{7.4}$$

As already discussed in Section 2, the reduced problem (RP$\mathcal{X}$) is ϵ-successful if the random subspace $\boldsymbol{p}+\text{range}(\boldsymbol{A})$ intersects the set of approximate global minimizers, which by Theorem 7.3 contains any connected components $\mathcal{G}^*_\epsilon$ defined in (7.3) for some global minimizer $\boldsymbol{x}^*$ of (P). Figure 4 shows an abstract representation of the situation where the random subspace $\boldsymbol{p} + \text{range}(\boldsymbol{A}_1)$ intersects the connected component $\mathcal{G}^*_\epsilon$, the corresponding embedding is therefore ϵ-successful; conversely, the random subspace $\boldsymbol{p} + \text{range}(\boldsymbol{A}_2)$ does not intersect $\mathcal{G}^*_\epsilon$. If $\mathcal{G}^*_\epsilon = G_\epsilon$ defined in (2.1), this implies that the corresponding embedding is not ϵ-successful.

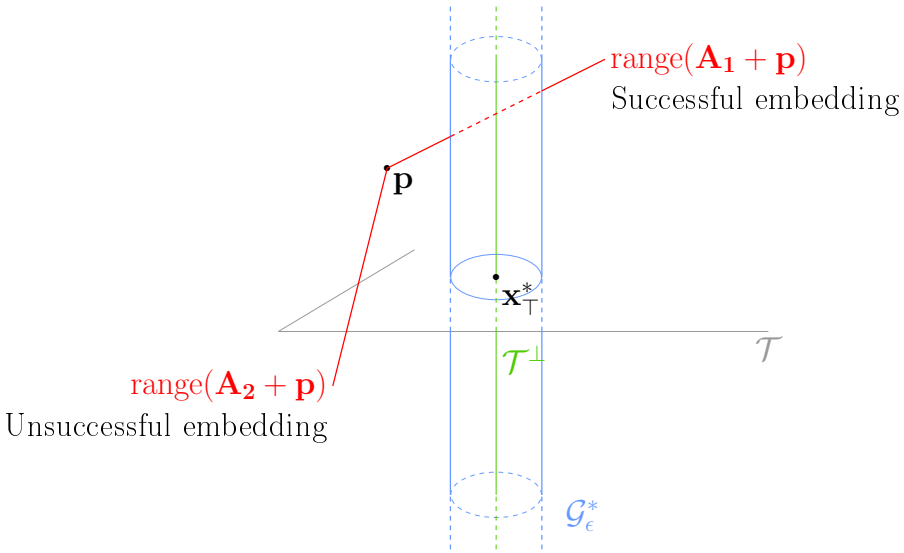

Figure 4: Abstract illustration of embeddings for functions with low effective dimension. The reduced problem is ϵ-successful if the random subspace intersects the connected component $\mathcal{G}^*_\epsilon$.

The following result further characterizes the probability of success of (RP$\mathcal{X}$).

Theorem 7.3. *Let $\mathcal{X} = \mathbb{R}^D$, and let Assumptions LipC and LowED hold. Let $\boldsymbol{A}$ be a $D \times d$ Gaussian matrix, $\boldsymbol{p} \in \mathbb{R}^D$ be a fixed vector, $\epsilon > 0$ an accuracy tolerance and $\boldsymbol{x}^*$ any global minimizer of* (P) *with associate connected component $\mathcal{G}^*_\epsilon$ as in* (7.3)*. Then,*

$$\begin{aligned}\mathbb{P}[(\text{RP}\mathcal{X}) \textit{ is } \epsilon\textit{-successful}] &\geq \mathbb{P}[\boldsymbol{p} + \text{range}(\boldsymbol{A}) \cap \mathcal{G}^*_\epsilon \neq \varnothing], \\ &= \mathbb{P}[\boldsymbol{U}^T\boldsymbol{p} + \text{range}(\boldsymbol{B}) \cap B_{\epsilon/L}(\boldsymbol{U}^T\boldsymbol{x}^*) \neq \varnothing],\end{aligned}$$

where $\boldsymbol{U}$ is an orthonormal matrix whose columns span the effective subspace $\mathcal{T}$ (see Assumption LowED), $\boldsymbol{B} := \boldsymbol{U}^T\boldsymbol{A}$, a $d_e \times d$ Gaussian matrix and $B_{\epsilon/L}(\boldsymbol{U}^T\boldsymbol{x}^*)$, the d_e-dimensional ball of radius ϵ/L centered at $\boldsymbol{U}^T\boldsymbol{x}^*$.

Proof. The first inequality simply follows from (2.2) and from the fact that $\mathcal{G}^*_\epsilon \subseteq G_\epsilon$, see (7.4). For the second relationship, since the matrix $\boldsymbol{Q} := \begin{bmatrix}\boldsymbol{U} & \boldsymbol{V}\end{bmatrix}$ (with $\boldsymbol{V}$ defined in Assumption LowED) is orthogonal, for all $\boldsymbol{y} \in \mathbb{R}^d$

$$\boldsymbol{A}\boldsymbol{y} + \boldsymbol{p} = \boldsymbol{Q}\boldsymbol{Q}^T(\boldsymbol{A}\boldsymbol{y} + \boldsymbol{p}) = \begin{bmatrix}\boldsymbol{U} & \boldsymbol{V}\end{bmatrix}\begin{bmatrix}\boldsymbol{U}^T \\ \boldsymbol{V}^T\end{bmatrix}(\boldsymbol{A}\boldsymbol{y} + \boldsymbol{p}) = (\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{V}\boldsymbol{V}^T)(\boldsymbol{A}\boldsymbol{y} + \boldsymbol{p}).$$

Writing $\boldsymbol{B} := \boldsymbol{U}^T\boldsymbol{A} \in \mathbb{R}^{d_e \times d}$ and $\boldsymbol{C} := \boldsymbol{V}^T\boldsymbol{A} \in \mathbb{R}^{(D-d_e)\times d}$, we get for any global minimizer $\boldsymbol{x}^*$ of (P) (with associate Euclidean projection $\boldsymbol{x}^*_\top$ on the effective subspace)

$$\begin{aligned}\boldsymbol{A}\boldsymbol{y} + \boldsymbol{p} &= \boldsymbol{U}(\boldsymbol{B}\boldsymbol{y} + \boldsymbol{U}^T\boldsymbol{p}) + \boldsymbol{V}(\boldsymbol{C}\boldsymbol{y} + \boldsymbol{V}^T\boldsymbol{p}) &(7.5)\\ &= \boldsymbol{x}^*_\top + \boldsymbol{U}(\boldsymbol{B}\boldsymbol{y} + \boldsymbol{U}^T\boldsymbol{p} - \boldsymbol{U}^T\boldsymbol{x}^*) + \boldsymbol{V}(\boldsymbol{C}\boldsymbol{y} + \boldsymbol{V}^T\boldsymbol{p}). &(7.6)\end{aligned}$$

By definition of $\mathcal{G}^*_\epsilon$, there follows that $\boldsymbol{A}\boldsymbol{y} + \boldsymbol{p} \in \mathcal{G}^*_\epsilon$ if and only if $\boldsymbol{B}\boldsymbol{y} + \boldsymbol{U}^T\boldsymbol{p} \in B_{\epsilon/L}(\boldsymbol{U}^T\boldsymbol{x}^*)$. By Theorem A.2, $\boldsymbol{B}$ is a $d_e \times d$ Gaussian matrix, which completes the proof. □

The probability of ϵ-success of (RP$\mathcal{X}$) can thus be lower bounded by the probability of the d-dimensional random subspace range$(\boldsymbol{B}) + \boldsymbol{U}^T\boldsymbol{p}$ intersecting the ball $B_{\epsilon/L}(\boldsymbol{U}^T\boldsymbol{x}^*)$ in $\mathbb{R}^{d_e}$. We now estimate the latter probability using the conic integral geometry results presented in Section 3 and Section 4: the two next results can be seen as the immediate counterparts of Theorem 4.3 and Theorem 4.4 for functions with low effective dimensionality.

Corollary 7.4. *Let $\mathcal{X} = \mathbb{R}^D$, and suppose that Assumptions LipC and LowED hold, with effective dimension $d_e > d$. Let $\epsilon > 0$ be an accuracy tolerance, $\boldsymbol{A}$, a $D \times d$ Gaussian matrix and $\boldsymbol{p} \in \mathbb{R}^D \setminus G_\epsilon$, a given vector. Let $r^{\text{eff}}_{\boldsymbol{p}} := \epsilon / \left(L\|\boldsymbol{U}^T\boldsymbol{x}^* - \boldsymbol{U}^T\boldsymbol{p}\|\right)$, where $\boldsymbol{x}^*$ is any global minimizer of* (P)*. Then*

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] \geq \tau(r^{\text{eff}}_{\boldsymbol{p}}, d, d_e), \tag{7.7}$$

where the function $\tau(r, d, d_e)$ for $0 < r < 1$ is defined in (4.5)*.*

Proof. The result is a direct extension of the analysis made in Section 4, and more precisely, Theorem 4.3, replacing $\boldsymbol{A}$ by $\boldsymbol{B}$, $\boldsymbol{x}^*$ by $\boldsymbol{U}^T\boldsymbol{x}^*$, $\boldsymbol{p}$ by $\boldsymbol{U}^T\boldsymbol{p}$ and D by d_e. □

Similarly as Theorem 4.4, the next result provides a uniform lower bound on the probability of ϵ-success of (RP$\mathcal{X}$).

Corollary 7.5. *Let $\mathcal{X} = \mathbb{R}^D$, and suppose that Assumptions LipC and LowED hold, with effective dimension $d_e > d$. Let $\epsilon > 0$ be an accuracy tolerance, $\boldsymbol{A}$, a $D \times d$ Gaussian matrix, $\boldsymbol{x}^*$, any global minimizer of* (P)*. Let $\boldsymbol{p} \in \mathbb{R}^D$ be a given vector that satisfies $\|\boldsymbol{U}^T\boldsymbol{p} - \boldsymbol{U}^T\boldsymbol{x}^*\| \leq R_{\max}$, for some suitably chosen $R_{\max}$, and let $r^{\text{eff}}_{\min} := \epsilon / (LR_{\max})$. Then*

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \epsilon\text{-successful}] \geq \tau(r^{\text{eff}}_{\min}, d, d_e), \tag{7.8}$$

where the function $\tau(r, d, d_e)$ for $0 < r < 1$ is defined in (4.5)*.*

Proof. The result is a mere adaptation of Theorem 4.4, replacing $\boldsymbol{A}$ by $\boldsymbol{B}$, $\boldsymbol{x}^*$ by $\boldsymbol{U}^T\boldsymbol{x}^*$, $\boldsymbol{p}$ by $\boldsymbol{U}^T\boldsymbol{p}$ and D by d_e. □

Note that adding some constraints (setting $\mathcal{X} \subset \mathbb{R}^D$) makes the analysis much more complicated as even if a random subspace $\{\boldsymbol{p} + \text{range}(\boldsymbol{A})\}$ intersects $\mathcal{G}^*_\epsilon$, this intersection may be outside the feasible domain; we therefore restrict ourselves to the unconstrained case in this paper.

7.3 X-REGO for functions with low effective dimension

We present an X-REGO variant dedicated to the optimization of functions with low effective dimension. This algorithm starts by exploring an embedding of low dimension d_{lb}, assuming $d_{lb} \leq d_e$, and the dimension is progressively increased until capturing the effective dimension of the problem, see Algorithm 2. Note that Line 3 to Line 6 are exactly the same as in Algorithm 1. Recall that Theorem 7.2 guarantees that the algorithm finds the global minimum of (P) with probability one if the reduced problem is solved exactly and if $d^k \geq d_e$, so that in this ideal

case we can terminate the algorithm after $d_e - d_{lb} + 1$ random embeddings; thus, Algorithm 2 terminates in finitely many random embeddings. Since the effective dimension is unknown, we typically terminate the algorithm when no progress is observed in the objective value, see Section 8 for numerical illustrations.

Algorithm 2 X-REGO for (P) when f has low effective dimension

1: Initialize $d^1 = d_{lb}$ for some $d_{lb} \geq 1$ and $\tilde{\boldsymbol{p}}^0 \in \mathcal{X}$
2: **for** $k \geq 1$ until termination **do**
3: Run lines 3 to 6 in Algorithm 1.
4: Let $d^{k+1} = d^k + 1$.
5: **end for**

7.4 Convergence of X-REGO for functions with low effective dimension

Similarly to Section 5 and Section 6, for each k, $\boldsymbol{p}^k$ is a random variable. The particular case of a deterministic $\boldsymbol{p}^k$ is represented using a random variable whose support is a singleton. To prove convergence of Algorithm 2 to an ϵ-minimizer while allowing the reduced problems to be solved approximately, we again require the reduced problems to be $(\epsilon - \lambda)$-successful, see Assumption Success-Solv and Assumption Succes-Emb. Note that unlike Section 6.2, the results below are finite termination results, as we known that with an ideal solver, Algorithm 2 finds an ϵ-minimizer after at most $d_e - d_{lb} + 1$ embeddings. Let us first show that Assumption Succes-Emb holds, and derive the value of τ_{lb}.

Corollary 7.6. *Suppose that $\mathcal{X} = \mathbb{R}^D$, that Assumptions LipC and LowED hold, that $\tilde{\boldsymbol{p}}^k$ satisfies $\|\boldsymbol{U}^T\tilde{\boldsymbol{p}}^k - \boldsymbol{U}^T\boldsymbol{x}^*\| \leq R_{\max}$ for all k and for some suitably chosen $R_{\max}$ and that $d_{lb} < d_e$. Then, Assumption Succes-Emb holds with*

$$\tau_{lb} = \tau(r_{\min}^{\text{eff}}, d_{lb}, d_e),$$

with $r_{\min}^{\text{eff}} = (\epsilon - \lambda)/(LR_{\max})$ and $\tau(\cdot,\cdot,\cdot)$ defined in (4.5).

Proof. For all embeddings such that $d^k < d_e$, the proof is the same as for Corollary 6.5, replacing D by d_e and $r_{\min}$ by $r_{\min}^{\text{eff}}$. Note that if $d^k \geq d_e$, (RP$\mathcal{X}$) is successful with probability one according to Theorem 7.2. The result follows then simply from the fact that $1 \geq \tau(r_{\min}, d_{lb}, d_e) = 2v_{d_e - d_{lb}+1}$ (see the Gauss-Bonnet formula (3.4), and the fact that the intrinsic volumes are nonnegative). □

The following result proves convergence of Algorithm 2 to the set of ϵ-minimizers almost surely after at most $d_e - d_{lb} + 1$ random embedding. Note in particular that this convergence result has no dependency on D.

Corollary 7.7 (Global convergence of X-REGO for functions with low effective dimension)**.** *Suppose that $\mathcal{X} = \mathbb{R}^D$, that Assumptions LipC, Success-Solv and LowED hold, and that $\tilde{\boldsymbol{p}}^k$ satisfies $\|\boldsymbol{U}^T\tilde{\boldsymbol{p}}^k - \boldsymbol{U}^T\boldsymbol{x}^*\| \leq R_{\max}$ for all k and for some suitably chosen $R_{\max}$. Suppose also that $d_{lb} \leq d_e$, let $\epsilon > 0$ be an accuracy tolerance, and let $k_{\max} = d_e - d_{lb} + 1$ be the index of the first embedding with dimension d_e. Then*

$$\mathbb{P}[f(\boldsymbol{x}_{opt}^{k_{\max}}) \leq f^* + \epsilon] > \rho^{k_{\max}}$$

where $\boldsymbol{x}_{opt}^k$ is defined in (5.2) and ρ^k is the probability of success of the solver for ($\widetilde{\text{RP}\mathcal{X}^k}$) (see Assumption Success-Solv). In particular, if the reduced problem is solved exactly, then $f(\boldsymbol{x}_{opt}^{k_{\max}}) \leq f^* + \epsilon$ with probability one. For $1 \leq k < k_{\max}$, we have

$$\mathbb{P}[f(\boldsymbol{x}_{opt}^k) \leq f^* + \epsilon] \geq 1 - (1 - \rho_{lb}\tau_{lb})^k$$

where $\tau_{lb} = \tau(r_{\min}^{\text{eff}}, d_{lb}, d_e)$*, with* $\tau(\cdot,\cdot,\cdot)$ *defined in* (4.5) *and* $r_{\min}^{\text{eff}} = (\epsilon - \lambda)/(LR_{\max})$.

Proof. Note that, by Corollary 7.6, Theorem 6.3 applies. However, since we are interested in finite termination results, we do not directly use Theorem 6.3; instead, we extract the following claim from its convergence proof, see (B.6). For all $K \geq 1$,

$$\mathbb{P}[\{\boldsymbol{x}_{opt}^K \in G_\epsilon\}] \geq 1 - \Pi_{k=1}^K (1 - \tau^k \rho^k). \tag{7.9}$$

It follows that

$$\mathbb{P}[\{\boldsymbol{x}_{opt}^K \in G_\epsilon\}] \geq 1 - (1 - \tau_{lb}\rho_{lb})^K, \tag{7.10}$$

where τ_{lb} and ρ_{lb} are computed/defined in Corollary 7.6 and Assumption Success-Solv, respectively. Finally, if $K \geq k_{\max}$, it follows that $d^K \geq d_e$, so that the probability of (RP$\mathcal{X}^k$) to be $(\epsilon - \lambda)$-successful is equal to one according to Theorem 7.2. So, if $K \geq k_{\max}$,

$$\mathbb{P}[\{\boldsymbol{x}_{opt}^K \in G_\epsilon\}] \geq 1 - (1 - \rho^K)\Pi_{k=1}^{K-1}(1 - \tau^k \rho^k) > 1 - (1 - \rho^K), \tag{7.11}$$

which concludes the proof.

□

8 Numerical experiments

Let us illustrate the behavior of X-REGO on a set of benchmark global optimization problems whose objectives have low effective dimension. We show empirically that Algorithm 2 simultaneously manages to accurately estimate the effective dimension of the problem, and outperforms significantly (and especially in the high-dimensional regime) the no-embedding framework, in which the original problem (P) is solved directly, with no exploitation of the special structure.

8.1 Setup

Test set. Our synthetic test set is very similar to the one we used in [15, 18], and contains a set of benchmark global optimization problems adapted to have low effective dimensionality in the objective as explained in Appendix C. Our test set is made of 18 D-dimensional functions with low effective dimension, with $D = 10, 100$ and 1000. These D-dimensional functions are constructed from 18 low-dimensional global optimization problems with known global minima (some of which are in the Dixon-Szego test set [23]), by artificially adding coordinates and then applying a rotation so that the effective subspace is not aligned with the coordinate axes.

Solver. The reduced problems are solved using the KNITRO solver ([13]). Note that, by default, KNITRO is a local solver, but switches to a global solver by activating its multistart feature. We therefore consider three "KNITRO"-type solvers: local KNITRO (no multistart used, referred to as KNITRO), and multistart KNITRO with a low/high number of starting points (cheap or expensive versions of multistart KNITRO, referred to here as ch-mKNITRO and exp-mKNITRO, respectively). The higher the number of starting points, the more likely the solver is to find a global minimizer of the reduced problem. See Table 1 for a detailed description of the settings of the different solvers.

Algorithms using a global solver (ch-mKNITRO and exp-mKNITRO). We test two different instances of the algorithmic framework presented in Algorithm 2 against the *no-embedding* framework, in which (P) is solved directly without using any random embedding and with no explicit exploitation of its special structure. For each instance, we let $d_{lb} = 1$. Since the effective dimension of the problem is assumed to be unknown, termination in Algorithm 2 is

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 ($\widetilde{\mathrm{RP}}\mathcal{X}^k$).

	exp-mKNITRO	ch-mKNITRO	KNITRO
Measure of computational cost	function evaluations	function evaluations	function evaluations
Max. budget to solve ($\widetilde{\mathrm{RP}}\mathcal{X}^k$)	$\min(200, 10d^k)$ starting points	$\min(100, 2d^k)$ starting points	1 starting point
Max. budget to solve (P)	$\min(200, 10D)$ starting points (only used for the *no-embedding* framework)	$\min(100, 2D)$ starting points (only used for the *no-embedding* framework)	1 starting point (only used for the *no-embedding* framework)
Termination for ($\widetilde{\mathrm{RP}}\mathcal{X}^k$) and (P)	Default options (unless overwritten by additional options)	Default options (unless overwritten by additional options)	Default options (unless overwritten by additional options)
Additional options for ($\widetilde{\mathrm{RP}}\mathcal{X}^k$) and (P)	`ms_enable` = 1, `ms_bndrange` = 2	`ms_enable` = 1, `ms_bndrange` = 2, `ms_maxsolves` = $\min(100, 2d^k)$ (for ($\widetilde{\mathrm{RP}}\mathcal{X}^k$)), `ms_maxsolves` = $\min(100, 2D)$ (for (P))	/

defined as the first embedding on which either stagnation is observed in the computed optimal cost of the reduced problem ($\widetilde{\mathrm{RP}}\mathcal{X}^k$), or if not, until $d^k = D$. Objective stagnation is measured as follows: stop after k_f embeddings, where k_f is the smallest $k \geq 2$ that satisfies

$$\left| f(\widetilde{\boldsymbol{A}}^k \tilde{\boldsymbol{y}}^k + \tilde{\boldsymbol{p}}^{k-1}) - f(\widetilde{\boldsymbol{A}}^{k-1} \tilde{\boldsymbol{y}}^{k-1} + \tilde{\boldsymbol{p}}^{k-2}) \right| \leq \gamma = 10^{-5}. \tag{8.1}$$

If $k_f \leq D$, we let $d_e^{\mathrm{est}} := k_f - 1$ be our estimate of the effective dimension of the problem. Indeed, by Theorem 7.2, two random problems of dimension d and $d+1$ with $d \geq d_e$ have the same optimal cost with probability one, so that the left-hand side of (8.1) would be zero if the reduced problems were solved exactly (i.e., under the assumption of an ideal solver). We argue that, on the other hand, it is very unlikely that two random reduced problems of dimension d and $d+1$ with $d < d_e$ have the same optimal cost.[12] We therefore terminate the algorithm after either $k = k_f$ (if there exists $k_f \leq D$ satisfying (8.1)), or else $k = D$ random embeddings. Regarding the choice of $\boldsymbol{p}^k$, we consider two possibilities: either $\boldsymbol{p}^k$ is a vector that does not depend on k, or $\boldsymbol{p}^k$ is the best point found over the k first embeddings (i.e., $\boldsymbol{p}^k = \boldsymbol{x}_{opt}^k$).

Algorithms relying on a local solver (KNITRO) and a resampling strategy. We also compare several instances of Algorithm 2 with the no-embedding framework when the reduced problem is solved using a local solver. Note that due to the possible nonconvexity of the problem, running a local solver on the original problem is not expected to find the global minimizer; results combining the no-embedding framework with a local solver are thus only reported for comparison. Recall also that our convergence analysis requires the solver to be able to find an approximate global minimizer of the subproblem with a sufficiently high probability. We show numerically that local solvers can be used when the points $\boldsymbol{p}^k$ are suitably chosen to globalize the search; we typically let $\boldsymbol{p}^k$, for some indices k, be a random variable with a sufficiently large support to contain a global minimizer of (P). Similarly as with global solvers, let k_f be the

[12] Admittedly, when optimizing difficult functions, for example that are flat almost everywhere and very steep around the minimizer, it could happen that two successive reduced problems of dimension d and $d+1$, with $d < d_e$, have the same optimal cost though none of them intersects the set of ϵ-minimizers; we exclude here such pathological cases.

smallest $k \geq 2$ that satisfies (8.1), and, if $k_f \leq D$, let $d_e^{\text{est}} := k_f - 1$ be our estimate of the effective dimension of the problem. However, since the solver is local, we cannot assume that (8.1) implies that we found an approximate global minimizer of the original problem (P). We therefore continue the optimization, fixing the subspace dimension: $d^k = d_e^{\text{est}}$ for all $k > k_f$, and assuming that $\boldsymbol{p}^k$ will be such that the next random subspace will leave the basin of attraction of the actual local minimizer. To prevent against local solutions, we use a stronger stopping criterion: the algorithm is stopped either after D embeddings, or earlier, when $k > k_f$ and when the computed optimal cost of the reduced problem did not change significantly over the last n_{stop} random embeddings, i.e., if

$$f(\boldsymbol{x}_{opt}^{k-n_{\text{stop}}+1}) - f(\boldsymbol{x}_{opt}^k) \leq \gamma = 10^{-5}. \tag{8.2}$$

In our experiments, we considered two possibilities: $n_{\text{stop}} = 3$ or $n_{\text{stop}} = 5$. Here again, we consider two main strategies for choosing $\boldsymbol{p}^k$: either $\boldsymbol{p}^k$ does not depend on k (e.g., $\boldsymbol{p}^k$ is an identically distributed random variable, for all k), or $\boldsymbol{p}^k$ is the best point found over the past embeddings ($\boldsymbol{p}^k = \boldsymbol{x}_{opt}^k$), resampling $\boldsymbol{p}^k$ at random in a sufficiently large domain for some values of k, see below.

Summary of the algorithms: In total, we compare four instances of Algorithm 2, that correspond to specific choices of $\boldsymbol{p}^k$, $k \geq 0$, and on the choice of a local/global solver.

- Adaptive X-REGO (A-REGO). In Algorithm 2, the reduced problem is solved using a global solver and the point $\boldsymbol{p}^k$ is chosen as the best point found[13] up to the kth embedding: $\boldsymbol{p}^k := \boldsymbol{A}^k\boldsymbol{y}^k + \boldsymbol{p}^{k-1}$.

- Local adaptive X-REGO (LA-REGO). In Algorithm 2, the reduced problem $(\widetilde{\text{RP}\mathcal{X}^k})$ is solved using a local solver (instead of global as in A-REGO). Until we find the effective dimension (i.e., for $k < k_f$), we use the same update rule for $\boldsymbol{p}^k$ as in A-REGO: $\boldsymbol{p}^k := \boldsymbol{A}^k\boldsymbol{y}^k + \boldsymbol{p}^{k-1}$. For the remaining embeddings, the point $\boldsymbol{p}^k$ is chosen as follows: $\boldsymbol{p}^k = \boldsymbol{A}^k\boldsymbol{y}^k + \boldsymbol{p}^{k-1}$ if $|f(\boldsymbol{A}^k\boldsymbol{y}^k + \boldsymbol{p}^{k-1}) - f(\boldsymbol{p}^{k-1})| > \gamma = 10^{-5}$, and $\boldsymbol{p}^k$ is draw uniformly in $[-1,1]^D$ otherwise, to compensate for the local behavior of the solver[14].

- Nonadaptive X-REGO (N-REGO). In Algorithm 2, the reduced problem is solved globally, and all the random subspaces are drawn at some fixed point: $\boldsymbol{p}^k = \boldsymbol{a}$. The fixed value $\boldsymbol{a}$ is simply defined as a realization of a random variable uniformly distributed in $[-1,1]^D$.[15]

- Local nonadaptive X-REGO (LN-REGO). In Algorithm 2, the reduced problem $(\widetilde{\text{RP}\mathcal{X}^k})$ is solved using a local solver. Until we find the effective dimension (i.e., for $k < k_f$), we set $\boldsymbol{p}^k = \boldsymbol{a}$, with $\boldsymbol{a}$ as in N-REGO. For $k \geq k_f$, $\boldsymbol{p}^k$ is a random variable distributed uniformly in $[-1,1]^D$ (and resampled at each embedding), to compensate for the local behavior of the solver.

Note that, regarding the choice of $\boldsymbol{p}^k$ when using a local solver, we typically have two phases. In the first phase, we apply the same selection rules for $\boldsymbol{p}^k$, $k < k_f$, as when using a global solver.

[13] If the reduced problem $(\widetilde{\text{RP}\mathcal{X}^k})$ is solved using a global solver, then $f(\tilde{\boldsymbol{p}}^k) \leq f(\tilde{\boldsymbol{p}}^{k-1})$ since $\tilde{\boldsymbol{p}}^{k-1}$ belongs to the search space of $(\widetilde{\text{RP}\mathcal{X}^k})$, so that we are indeed keeping the best point found so far. If we are using a local solver, we always initialize $\boldsymbol{y} = \boldsymbol{0}$ when solving $(\widetilde{\text{RP}\mathcal{X}^k})$, so that the same conclusion holds.

[14] We know, from the way we have constructed the test set, that for each problem there exists a global minimizer that belongs to $[-1,1]^D$.

[15] One could take simply $\boldsymbol{p} = \boldsymbol{0}$, but due to the way we have constructed the problem set, setting $\boldsymbol{p} = \boldsymbol{0}$ may give some advantage to the algorithm, so we let $\boldsymbol{p}^k = \boldsymbol{a}$ instead, where $\boldsymbol{a}$ is a random variable drawn once at the beginning of the algorithm.

For $k \geq k_f$, we allow resampling to avoid the algorithm to be trapped at a local minimizer. We do not introduce some resampling in the first phase, because then stochasticity would impact the criterion (8.1) and our estimate of the effective dimension of the problem.

Experimental setup. For each algorithm described above, we solve the entire test set 3 times to estimate the average performance of the algorithms, and record the computational cost, which we measure in terms of function evaluations (the termination criterion is described above). Note that from the four algorithms described above, we get six different algorithms, since algorithms A-REGO and N-REGO are endowed with two different global solvers: exp-mKNITRO and ch-mKNITRO, corresponding respectively to a low and large number of starting points. To compare with 'no-embedding', we solve the full-dimensional problem (P) directly with the corresponding solver with no use of random embeddings. The budget and termination criteria used to solve $(\widetilde{\mathrm{RP}\mathcal{X}^k})$ within X-REGO or to solve (P) in the 'no-embedding' framework are the default ones, summarized in Table 1.

Remark 8.1. 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 present the main numerical results using Dolan and Moré's performance profile [25] — a popular framework to compare the performance of optimization algorithms applied to a given test set. For a given algorithm $\mathcal{A}$, and for each function f in the test set $\mathcal{S}$, we define

$$N_f(\mathcal{A}) := \text{min. \# of fun. evals required by the algorithm to converge.}$$

If $\mathcal{A}$ fails to successfully converge to a ϵ-minimizer of f, with $\epsilon = 10^{-3}$, within the maximum computational budget, we set $N_f(\mathcal{A}) = \infty$. We further define

$$N_f^* := \min_{\mathcal{A}} N_f(\mathcal{A}),$$

as the minimal computational cost required by any algorithm to optimize f. We normalize all the computational costs by N_f^* and, for each $\mathcal{A}$, we plot a function $\pi_{\mathcal{A}}(\alpha)$ that computes the proportion of f's in the test set $\mathcal{S}$, for which the normalized computational effort spent by $\mathcal{A}$ was less than α. Mathematically speaking,

$$\pi_{\mathcal{A}}(\alpha) := \frac{|\{f : N_f(\mathcal{A}) \leq \alpha N_f^*\}|}{|\mathcal{S}|} \text{ for } \alpha \geq 1,$$

where $|\cdot|$ denotes the cardinality of a set. The algorithm $\mathcal{A}$ is considered to have achieved better performance if it produces higher values for $\pi_{\mathcal{A}}(\alpha)$ for lower values of α, i.e., on figures, the curve $\pi_{\mathcal{A}}(\alpha)$ is higher and lefter.

8.2 Numerical results

Comparison of X-REGO with the no-embedding framework. The comparison between the above-mentioned instances of X-REGO and the no-embedding framework is given in Figure 5. A-REGO and N-REGO clearly outperform the no-embedding framework in terms of accuracy vs computational cost, especially for large D. Reducing the number of starting points in the multistart strategy (i.e., replacing exp-mKNITRO by ch-mKNITRO) allows to further significantly improve the performance, though the total proportion of problems ultimately solved is slightly decreased compared to exp-mKNITRO. Note also that the use of a local solver (LA-REGO and LN-REGO) outperforms both global X-REGO instances and the no-embedding framework, especially for large D. They find the global minimizer in a significantly higher number of subproblems than when directly addressing the original high-dimensional problem with

the local solver: the resampling strategy for $\boldsymbol{p}^k$ described above helps to globalize the search. Table 2 contains the average, over the test problems, of the number of embeddings used per algorithm; note that for (approximately) global solvers, and especially using $\boldsymbol{p}^k = \boldsymbol{x}^k_{opt}$, the average number of embeddings is very close to the ideal k_f. Indeed, the average effective dimension on our problem sets is equal to 3.7, so the ideal average number of embeddings should be 4.7, as we need an additional embedding for the stopping criterion (8.1) to be satisfied. For local solvers, the average number of embeddings is slightly higher due to the need to resample candidate solutions to globalize the search and due to the stronger stopping criterion.

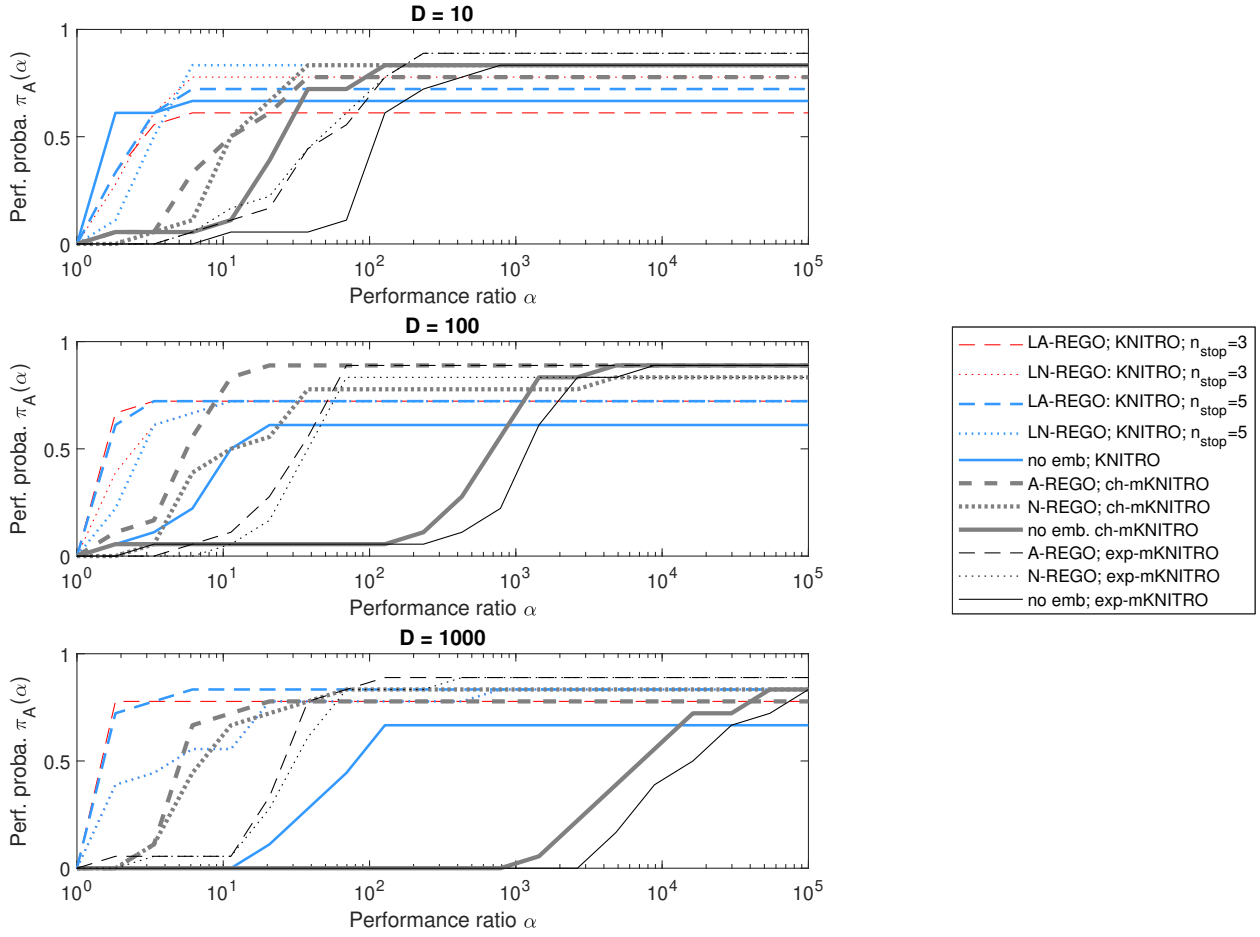

Figure 5: Comparison between the X-REGO algorithms and 'no-embedding' with KNITRO. Each algorithm was run three times on the whole dataset; since all three runs returned similar curves, we display only one of them.

Table 2: Average number of embeddings per problem, estimated from 3 independent runs of the algorithms on the test set.

	KNITRO ($n_{\text{stop}} = 3$)		KNITRO ($n_{\text{stop}} = 5$)		ch-mKNITRO		exp-mKNITRO	
	LA-REGO	LN-REGO	LA-REGO	LN-REGO	A-REGO	N-REGO	A-REGO	N-REGO
$D = 10$	5.96	6.59	7.87	8.07	4.78	5.19	4.70	4.89
$D = 100$	6.41	7.81	9.09	9.33	4.76	8.94	4.70	5.63
$D = 1000$	6.26	9.15	8.54	10.83	4.67	7.17	4.67	6.67

Estimation of the effective dimension. As described earlier, instances of X-REGO naturally provide an estimate d_e^{est} of the effective dimension of the problem: $d_e^{\text{est}} = k_f - 1$, where k_f is the smallest integer that satisfies (8.1). In case there exists no $k_f \leq D$ satisfying (8.1), we set $d_e^{\text{est}} = D$. For several instances of Algorithm 2, Table 3 reports the number of problems of the data set on which $d_e^{\text{est}} \in [d_e, d_e + 2]$, where d_e is the exact effective dimension of the problem, for $D = 10$, $D = 100$ and $D = 1000$. Typically, adaptive choices of $\boldsymbol{p}^k$ results in a slightly larger estimate of the effective dimension; we also note that the use of a local solver is comparable to a global one regarding the ability of the algorithm to estimate the effective dimension on this problem set when $\boldsymbol{p}^k$ is chosen adaptively, and significantly lower otherwise. The values given in Table 3 have been averaged over three independent runs of our experiment, on the whole dataset, to account for randomness in the algorithms.

Table 3: Percentage of problems for which the estimated effective dimension lies in the interval $[d_e, d_e + 2]$, where d_e is the true effective dimension of the problem.

	LA-REGO KNITRO	LN-REGO KNITRO	A-REGO ch-mKNITRO	N-REGO ch-mKNITRO	A-REGO exp-mKNITRO	N-REGO exp-mKNITRO
$D = 10$	94.44	79.63	94.44	83.33	94.44	88.89
$D = 100$	94.44	68.52	94.44	79.63	94.44	85.19
$D = 1000$	94.44	66.67	88.89	81.48	90.74	85.19

What if we know the effective dimension of the problem? In the favorable situation when the effective dimension d_e of each problem is known, we can set $d_{lb} = d_e$ in Algorithm 2, and theoretically, for an ideal global solver, Algorithm 2 is guaranteed to solve exactly the original problem using one embedding. Figure 6 explores numerically the validity of this claim. We compare several instances of X-REGO with corresponding counterparts, where the effective dimension is known. When using an (approximately) global solver (ch-mKNITRO or exp-mKNITRO), we stop Algorithm 2 after one embedding of dimension d_e. When the solver is local (KNITRO), we let Algorithm 2 explore several embeddings of dimension d_e, and stop the algorithm when (8.2) is satisfied, with $n_{\text{stop}} = 3$, or otherwise after 50 embeddings. Figure 6 shows the corresponding performance profiles, when comparing these strategies with the ones presented on Figure 4, and the corresponding no-embedding algorithms. In general, and except when using local solvers, knowing d_e allows to solve a significant proportion of the problems in a considerably smaller time. Admittedly, these conclusions strongly depend on the probability of the solver to be successful, i.e., of the number of starting points of the multistart procedure. Note also than in our test set, the effective dimension is typically low (average value is 3.7), which might also decrease the benefit of knowing the effective dimension and thus avoiding to explore lower-dimensional subspaces; we expect the gap between Algorithm 2 and algorithms where d_e is known to increase with the effective dimension of the problem.

8.3 Conclusions to numerical experiments.

We have compared several instances of Algorithm 2 with the no-embedding framework, where the original problem is addressed directly, with no use of random embeddings nor exploitation of the special structure. Overall, Algorithm 2 outperforms the no-embedding framework, and this observation becomes more and more apparent when the dimension of the original problem increases. We have also combined Algorithm 2 with a local solver; though our convergence theory does not cover this situation, we have shown that the resulting algorithm can outperform both the no-embedding framework and instances of Algorithm 2 relying on global solver when the parameters $\boldsymbol{p}^k$ are sampled at random in a sufficiently large domain to "globalize" the search. Regarding the estimation of the effective dimension, we noticed that instances of Algorithm 2

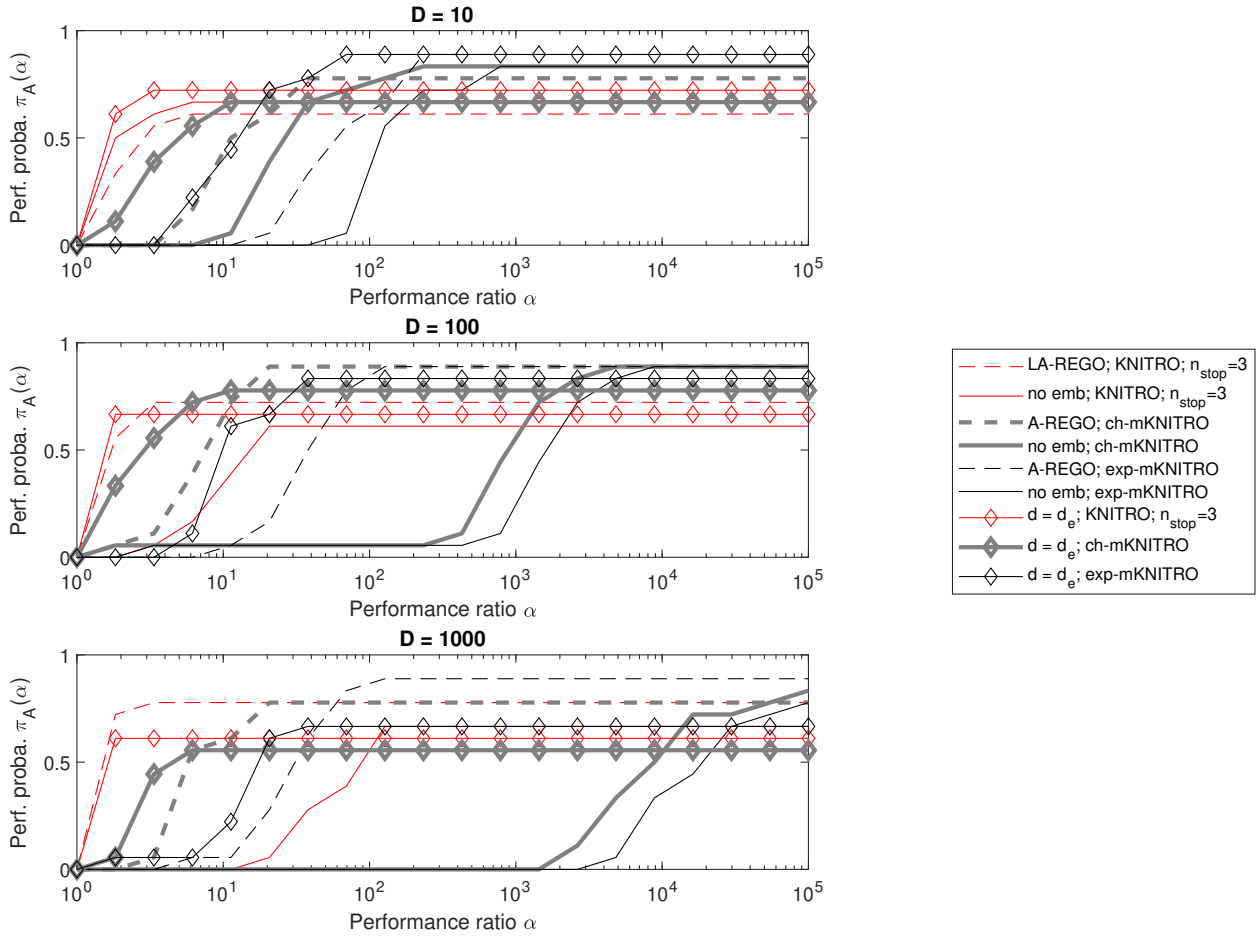

Figure 6: Comparison of X-REGO with no-embedding, and variants of X-REGO in which the random subspace dimension is equal to the effective dimension (assumed to be known).

relying on adaptive rules for selecting $\boldsymbol{p}^k$ (A-REGO and LA-REGO) significantly outperform their fixed $\boldsymbol{p}^k$ counterparts. Finally, we have shown that, in the favourable case when the effective dimension is known, letting $d_{lb} \geq d_e$ in Algorithm 2 leads to a substantial improvement in performance.

9 Conclusions and future work

We explored a generic algorithmic framework, X-REGO, for global optimization of Lipschitz-continuous functions. X-REGO is based on successively generating reduced problems (RP$\mathcal{X}$), where the parameter $\boldsymbol{p}$ is flexibly-chosen. Flexibility in choosing $\boldsymbol{p}$ allows the user to calibrate the level of exploration in $\mathcal{X}$.

Our central result is the proof of global convergence of X-REGO, that heavily relies on an estimate of the probability of the reduced problem (RP$\mathcal{X}$) to be ϵ-successful. By looking at the reduced problem through the prism of conic geometry, we have developed a new type of analysis to bound the probability of ϵ-success of (RP$\mathcal{X}$). The bounds are expressed in terms of the so-called conic intrinsic volumes of circular cones which have exact formulae and thus are quantifiable. Using these formulae, we analysed the asymptotic behaviour of the bounds for large D. The analysis suggests that the success rate of (RP$\mathcal{X}$) (as expected) decreases exponentially with growing D. Confirming our intuition, the analysis also shows that (RP$\mathcal{X}$)

has a high success rate for larger d and smaller distances between the location where subspaces are embedded (i.e., the point $\boldsymbol{p}$) and the location of a global minimizer $\boldsymbol{x}^*$. This latter property of (RP$\mathcal{X}$) for general Lipschitz continuous functions is remindful of the dependence of the success rates of (RP$\mathcal{X}$) for functions with low effective dimensionality on the distance between $\boldsymbol{p}_\top$ and $\boldsymbol{x}^*_\top$, see [18]. Furthermore, to understand the relative performance of (RP$\mathcal{X}$), we compared it with a uniform sampling technique. We looked at lower bounds for the probability of ϵ-success of the two techniques and found that the lower bound $\tau(r_{\boldsymbol{p}}, d, D)$ for (RP$\mathcal{X}$) is greater than the lower bound τ_{us} for uniform sampling if the distance $\|\boldsymbol{x}^* - \boldsymbol{p}\|$ is smaller than $0.48\sqrt{D}$ in the asymptotic regime ($D \to \infty$). In the asymptotic analysis, the embedding subspace d was kept fixed. The analysis showed that in this regime d has no significant effect on the relative performance of (RP$\mathcal{X}$). Future research may involve comparison of the performances of (RP$\mathcal{X}$) and uniform sampling in different asymptotic settings, for example, when $d = \beta D$ for some fixed constant β.

Our derivations are conceptual in nature, exploring new connections of global optimization to other areas such as conic integral geometry. As an illustration, in the second part of the paper, we used our analysis to obtain lower bounds — that are independent of D — for the probability of ϵ-success of (RP$\mathcal{X}$) for functions with low effective dimensionality in the case $d < d_e$. This analysis is exploited algorithmically and allows lifting the restriction of needing to know d_e for random embeddings algorithms for functions with low effective dimensionality. We tested the effectiveness of X-REGO numerically using global and local KNITRO for solving the reduced problem on a set of benchmark global optimization problems modified to have low effective dimensionality. We proposed different variants of X-REGO each corresponding to a specific rule for choosing $\boldsymbol{p}$'s and contrasted them against each other and against the 'no-embedding' framework in which the solvers were applied to (P) directly with no use of subspace embeddings. The results of the experiments showed that the difference in performance between X-REGO and 'no-embedding' becomes more prominent for larger D, in favour of X-REGO. The results further suggest that the effectiveness of X-REGO, just like of REGO in [15], is solver-dependent. In our experiments, the best results were achieved by the local solver. In the future, we plan to investigate the performance of X-REGO when applied to general objectives and compare it with popular global optimization solvers.

References

[1] D. Amelunxen. *Geometric analysis of the condition of the convex feasibility problem.* PhD thesis, University of Paderborn, 2011.

[2] D. Amelunxen and M. Lotz. Intrinsic volumes of polyhedral cones: A combinatorial perspective. *Discrete & Computational Geometry*, 58(2):371–409, 2017.

[3] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

[4] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. *Optimization with Sparsity-Inducing Penalties*, volume 4:1. Foundations and Trends in Machine Learning, 2011.

[5] 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.

[6] A. Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB.* MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, 2014.

[7] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization.* Society for Industrial and Applied Mathematics, 2001.

[8] A. S. Berahas, R. Bollapragada, and J. Nocedal. An investigation of Newton-sketch and subsampled Newton methods. *Optimization Methods and Software*, 35(4):661–680, 2020.

[9] E. H. Bergou, E. Gorbunov, and P. Richtárik. Stochastic three points method for unconstrained smooth minimization. *SIAM J. Optim.*, 30(4):2726–2749, 2020.

[10] M. Binois, D. Ginsbourger, and O. Roustant. A warped kernel improving robustness in Bayesian optimization via random embeddings. *arXiv e-prints*, page arXiv:1411.3685, 2014.

[11] M. Binois, D. Ginsbourger, and O. Roustant. On the choice of the low-dimensional domain for global optimization via random embeddings. *arXiv e-prints*, page arXiv:1704.05318, 2017.

[12] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

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

[14] H. Cai, D. Mckenzie, W. Yin, and Z. Zhang. Zeroth-Order Regularized Optimization (ZORO): Approximately Sparse Gradients and Adaptive Sampling. *arXiv e-prints*, page arXiv:2003.13001, 2020.

[15] 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.

[16] C. Cartis and L. Roberts. Scalable subspace methods for derivative-free nonlinear least-squares optimization. *arXiv e-prints*, page arxiv:2102.12016, 2021.

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

[18] C. Cartis, E. Massart, and A. Otemissov. Constrained global optimization of functions with low effective dimensionality using random subspaces. *arXiv e-prints*, page arXiv:2009.10446, 2020.

[19] C. Cartis, J. Fiala, and Z. Shao. Hashing embeddings of optimal dimension, with applications to linear least squares. *arXiv e-prints*, page arxiv:2105.11815, 2021.

[20] 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.

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

[22] 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.

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

[24] 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.

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

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

[27] 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.

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

[29] 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.

[30] 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.

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

[32] M. Gendreau and J.-Y. Potvin. *Handbook of Metaheuristics*. International Series in Operations Research & Management Science. Springer US, 2nd edition, 2010.

[33] L. Goldstein, I. Nourdin, and G. Peccati. Gaussian phase transitions and conic intrinsic volumes: Steining the steiner formula. *The Annals of Applied Probability*, 27(1):1–47, 2017.

[34] D. Golovin, J. Karro, G. Kochanski, C. Lee, X. Song, and Q. Zhang. Gradientless descent: High-dimensional zeroth-order optimization. In *Proceedings of the Sixth International Conference on Learning Representations*, ICLR'20, 2020.

[35] 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. 2019.

[36] S. Gratton, C. W. Royer, L. N. Vicente, and Z. Zhang. Direct search based on probabilistic descent. *SIAM J. Optim.*, 25(3):1515–1541, 2015.

[37] Serge Gratton, Clément W. Royer, Luís N. Vicente, and Zaikun Zhang. Direct search based on probabilistic feasible descent for bound and linearly constrained problems. *Computational Optimization and Applications*, 72(3):525–559, 2019.

[38] D. Grishchenko, F. Iutzeler, and J. Malick. Proximal gradient methods with adaptive subspace sampling. *Mathematics of Operations Research*, 2021.

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

[40] F. Hanzely, N. Doikov, P. Richtárik, and Y. Nesterov. Stochastic subspace cubic Newton method. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, 2020.

[41] J. H. Holland. Genetic algorithms and the optimal allocation of trials. *SIAM Journal on Computing*, 2(2): 88–105, 1973.

[42] P. Izmailov, W. J. Maddox, P. Kirichenko, T. Garipov, D. Vetrov, and A. G. Wilson. Subspace inference for Bayesian deep learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI'19)*, 2019.

[43] 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*, ICML'19, 2019.

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

[45] J. Lacotte and M. Pilanci. Effective dimension adaptive sketching methods for faster regularized least-squares optimization. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, NIPS'20, 2020.

[46] J. Lacotte, M. Pilanci, and M. Pavone. High-dimensional optimization in adaptive random subspaces. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, NIPS'19, 2019.

[47] C. Li, H. Farkhoor, R. Liu, and J. Yosinski. Measuring the intrinsic dimension of objective landscapes. In *Proceedings of the Sixth International Conference on Learning Representations*, ICLR'18, 2018.

[48] M. B. McCoy and J. A. Tropp. From Steiner formulas for cones to concentration of intrinsic volumes. *Discrete & Computational Geometry*, 51(4):926–963, 2014.

[49] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

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

[51] NIST. NIST Digital Library of Mathematical Functions, 2020. Available at https://dlmf.nist.gov.

[52] A. Otemissov. *Dimensionality Reduction Techniques for Global Optimization*. PhD thesis, University of Oxford, 2021.

[53] M. Pilanci and M. J. Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

[54] 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, 2016.

[55] P. Richtárik and M. Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156:433–484, 2015.

[56] F. Roosta-Khorasani and M. W. Mahoney. Sub-sampled Newton methods. *Mathematical Programming*, 174 (1):293–326, 2019.

[57] R. Schneider and W.Weil. *Stochastic and Integral Geometry*. Springer series in statistics: Probability and its applications. Springer, 2008.

[58] F. J. Solis and R. J.-B. Wets. Minimization by random search techniques. *Mathematics of Operations Research*, 6(1), 1981.

[59] S. U. Stich, C. L. Müller, and B. Gärtner. Optimization of convex functions with random pursuit. *SIAM Journal on Optimization*, 23(2):1284–1309, 2013.

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

[61] F. G. Tricomi and A. Erdélyi. The asymptotic expansion of a ratio of gamma functions. *Pacific J. Math.*, 1: 133–142, 1951.

[62] 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.

[63] G. Ughi, V. Abrol, and J. Tanner. An empirical study of derivative-free-optimization algorithms for targeted black-box attacks in deep neural networks. *Optimization and Engineering*, 2021.

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

[65] Y. Wang, S. S. Du, S. Balakrishnan, and A. Singh. Stochastic zeroth-order optimization in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, AISTATS'18, 2018.

[66] 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.

[67] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[68] S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151:3—34, 2015.

[69] 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, 2019.

A Technical definitions and results

A.1 Gaussian random matrices

Definition A.1 (Gaussian matrix)**.** A Gaussian (random) matrix is a matrix whose each entry is an independent standard normal random variable.

Gaussian matrices have been well-studied with many results available at hand; here, we mention the following result that we use in the analysis; for a collection of results pertaining to Gaussian matrices and other related distributions refer to [39, 64].

Theorem A.2. *(see [39, 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.*

B Global convergence proof

This section contains material already presented in [18], with minor changes to capture the fact that the lower bounds ρ^k and τ^k are now variable with k (or, in other words, the probability that the reduced problem $(\widetilde{\mathrm{RP}\mathcal{X}^k})$ is ϵ-successful, as well as the probability that the solver finds a sufficiently accurate solution of the reduced problem, is changing with the dimension of the reduced problem d^k in Algorithm 1). The following three lemmas are needed in our convergence proof.

Lemma B.1. *If Assumption Success-Solv holds, then*

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

Proof. Assumption Success-Solv implies

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

where the equality follows from the fact that R^k is $\mathcal{F}^{k-1/2}$-measurable and, thus, can be pulled out of the expectation (see [26, Theorem 4.1.14]). □

A useful property is given next.

Lemma B.2. *Let Assumption Success-Solv and Succes-Emb hold. Then, for $K \geq 1$, we have*

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

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

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

where

- $(*)$ follow from the tower property of conditional expectation (see (4.1.5) in [26]),
- $(\circ)$ 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 [26]),
- the inequalities follow from (B.1) and (6.5), 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.

Lemma B.3. *Suppose Assumption Success-Solv and Succes-Emb hold. Then,*

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

Proof. By Definition 1.1, if (RP$\mathcal{X}^k$) is $(\epsilon-\lambda)$-successful, then there exists $\boldsymbol{y}^k_{int} \in \mathbb{R}^{d^k}$ such that $\boldsymbol{A}^k\boldsymbol{y}^k_{int} + \boldsymbol{p}^{k-1} \in \mathcal{X}$ and

$$f(\boldsymbol{A}^k\boldsymbol{y}^k_{int} + \boldsymbol{p}^{k-1}) \leq f^* + \epsilon - \lambda. \tag{B.2}$$

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

$$f^k_{min} \leq f(\boldsymbol{A}^k\boldsymbol{y}^k_{int} + \boldsymbol{p}^{k-1}). \tag{B.3}$$

Then, for $\boldsymbol{x}^k$, (6.2) 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 (B.3) and (B.2), respectively. This shows that $\boldsymbol{x}^k \in G_\epsilon$. □

B.1 Proof of Theorem 6.3.

Lemma B.3 and the definition of $\boldsymbol{x}^k_{opt}$ in (5.2) 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,

$$\bigcup_{k=1}^{K} \{R^k = 1\} \cap \{S^k = 1\} \subseteq \bigcup_{k=1}^{K} \{\boldsymbol{x}^k_{opt} \in G_\epsilon\}. \tag{B.4}$$

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,

$$\bigcup_{k=1}^{K} \{\boldsymbol{x}^k_{opt} \in G_\epsilon\} = \{\boldsymbol{x}^K_{opt} \in G_\epsilon\}. \tag{B.5}$$

From (B.4) and (B.5), we have for all $K \geq 1$,

$$\mathbb{P}[\{\boldsymbol{x}^K_{opt} \in G_\epsilon\}] \geq \mathbb{P}\Big[\bigcup_{k=1}^{K} \{R^k = 1\} \cap \{S^k = 1\}\Big] \geq 1 - \Pi_{k=1}^K (1 - \tau^k\rho^k), \tag{B.6}$$

where the second inequality follows from Lemma B.2. Finally, passing to the limit with K in (B.6), we deduce

$$1 \geq \lim_{K\to\infty} \mathbb{P}[\{\boldsymbol{x}_{opt}^K \in G_\epsilon\}] \geq \lim_{K\to\infty} \left[1 - \Pi_{k=1}^K (1 - \tau^k \rho^k)\right] \geq \lim_{K\to\infty} \left[1 - (1 - \tau_{lb}\rho_{lb})^K\right] = 1,$$

with τ_{lb} and ρ_{lb} defined in Assumption Succes-Emb and Assumption Success-Solv, respectively. Since $\tau_{lb}\rho_{lb} > 0$ by Assumption Success-Solv and Assumption Succes-Emb, we get the required result. Note that if

$$1 - (1 - \tau_{lb}\rho_{lb})^k \geq \xi \tag{B.7}$$

then (B.6) implies $\mathbb{P}[\boldsymbol{x}_{opt}^k \in G_\epsilon] \geq \xi$. Since (B.7) is equivalent to $k \geq \dfrac{\log(1-\xi)}{\log(1-\tau_{lb}\rho_{lb})}$, (B.7) holds for all $k \geq K_\xi$ since $K_\xi \geq \dfrac{\log(1-\xi)}{\log(1-\tau_{lb}\rho_{lb})}$.

C Problem set

Table 4 contains the name, domain and global minimum of the functions used to generate the high-dimensional test set. Similarly as in [15, 18], the problem set contains 18 problems taken from [31, 28, 60]. To generate this problem set, we transformed each of the 18 functions in Table 4 into a high-dimensional function with low-effective dimension, by adapting the method proposed by Wang et al. [66]. Let $\bar{g}(\bar{\boldsymbol{x}})$ be any function from Table 4, with dimension d_e 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 nontrivial constant subspace. The final form of the function we test is

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

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 according to (C.1), one for each $D = 10, 100, 1000$. Note that the range of effective dimension covered by our test set is slightly larger than in [15, 18], to better assess the ability of the algorithm to learn d_e.

Table 4: The problem set listed in alphabetical order.

Function	Domain	Global minima
1) Beale [28]	$\boldsymbol{x} \in [-4.5, 4.5]^2$	$g(\boldsymbol{x}^*) = 0$
2) Branin [28]	$x_1 \in [-5, 10]$ $x_2 \in [0, 15]$	$g(\boldsymbol{x}^*) = 0.397887$
3) Brent [31]	$\boldsymbol{x} \in [-10, 10]^2$	$g(\boldsymbol{x}^*) = 0$
5) Easom [28]	$\boldsymbol{x} \in [-100, 100]^2$	$g(\boldsymbol{x}^*) = -1$
6) Goldstein-Price [28]	$\boldsymbol{x} \in [-2, 2]^2$	$g(\boldsymbol{x}^*) = 3$
7) Hartmann 3 [28]	$\boldsymbol{x} \in [0, 1]^3$	$g(\boldsymbol{x}^*) = -3.86278$
8) Hartmann 6 [28]	$\boldsymbol{x} \in [0, 1]^6$	$g(\boldsymbol{x}^*) = -3.32237$
9) Levy [60]	$\boldsymbol{x} \in [-10, 10]^6$	$g(\boldsymbol{x}^*) = 0$
10) Perm 4, 0.5 [60]	$\boldsymbol{x} \in [-4, 4]^4$	$g(\boldsymbol{x}^*) = 0$
11) Rosenbrock [60]	$\boldsymbol{x} \in [-5, 10]^7$	$g(\boldsymbol{x}^*) = 0$
12) Shekel 5 [60]	$\boldsymbol{x} \in [0, 10]^4$	$g(\boldsymbol{x}^*) = -10.1532$
13) Shekel 7 [60]	$\boldsymbol{x} \in [0, 10]^4$	$g(\boldsymbol{x}^*) = -10.4029$
14) Shekel 10 [60]	$\boldsymbol{x} \in [0, 10]^4$	$g(\boldsymbol{x}^*) = -10.5364$
15) Shubert [60]	$\boldsymbol{x} \in [-10, 10]^2$	$g(\boldsymbol{x}^*) = -186.7309$
16) Six-hump camel [60]	$x_1 \in [-3, 3]$ $x_2 \in [-2, 2]$	$g(\boldsymbol{x}^*) = -1.0316$
17) Styblinski-Tang [60]	$\boldsymbol{x} \in [-5, 5]^8$	$g(\boldsymbol{x}^*) = -313.329$
18) Trid [60]	$\boldsymbol{x} \in [-25, 25]^5$	$g(\boldsymbol{x}^*) = -30$
19) Zettl [28]	$\boldsymbol{x} \in [-5, 5]^2$	$g(\boldsymbol{x}^*) = -0.00379$