On the choice of the low-dimensional domain for global optimization via random embeddings

Mickaël Binois, David Ginsbourger, Olivier Roustant

On the choice of the low-dimensional domain for global optimization via random embeddings

Mickaël Binois* David Ginsbourger †‡ Olivier Roustant§

October 3, 2018

Abstract

The challenge of taking many variables into account in optimization problems may be overcome under the hypothesis of low effective dimensionality. Then, the search of solutions can be reduced to the random embedding of a low dimensional space into the original one, resulting in a more manageable optimization problem. Specifically, in the case of time consuming black-box functions and when the budget of evaluations is severely limited, global optimization with random embeddings appears as a sound alternative to random search. Yet, in the case of box constraints on the native variables, defining suitable bounds on a low dimensional domain appears to be complex. Indeed, a small search domain does not guarantee to find a solution even under restrictive hypotheses about the function, while a larger one may slow down convergence dramatically. Here we tackle the issue of low-dimensional domain selection based on a detailed study of the properties of the random embedding, giving insight on the aforementioned difficulties. In particular, we describe a minimal low-dimensional set in correspondence with the embedded search space. We additionally show that an alternative equivalent embedding procedure yields simultaneously a simpler definition of the low-dimensional minimal set and better properties in practice. Finally, the performance and robustness gains of the proposed enhancements for Bayesian optimization are illustrated on numerical examples.

Keywords: Expensive black-box optimization; low effective dimensionality; zonotope; REMBO; Bayesian optimization

1 Introduction

Dealing with many variables in global optimization problems has a dramatic impact on the search of solutions, along with increased computational times, see e.g., [45]. This effect is

*Corresponding author: The University of Chicago Booth School of Business, 5807 S. Woodlawn Ave., Chicago, IL, USA; mbinois@chicagobooth.edu

†Uncertainty Quantification and Optimal Design group, Idiap Research Institute, Centre du Parc, Rue Marconi 19, PO Box 592, 1920 Martigny, Switzerland

‡IMSV, Department of Mathematics and Statistics, University of Bern, Alpeneggstrasse 22, 3012 Bern, Switzerland

§EMSE Ecole des Mines de St-Étienne, UMR CNRS 6158, LIMOS, F-42023: 158 Cours Fauriel, Saint-Étienne, France

particularly severe for methods dedicated to black-box, derivative free expensive-to-evaluate problems. The latter are crucial in engineering, and generally in all disciplines calling for complex mathematical models whose analysis relies on intensive numerical simulations. Among dedicated methods, those from Bayesian Optimization (BO) rely on a surrogate model to save evaluations, such as Gaussian Processes (GPs). They have known a fantastic development in the last two decades, both in the engineering and machine learning communities, see e.g., [47]. Successful extensions include dealing with stochasticity [22], variable fidelity levels [8] as well as with constrained and multi-objective setups [15].

Now, standard implementation of such algorithms are typically limited in terms of dimensionality because of the type of covariance kernels often used in practice. The root of the difficulty with many variables is that the number of observations required to learn a function without additional restrictive assumptions increases exponentially with the dimension, which is known as the "curse of dimensionality", see e.g., [11], [20]. Here, we consider the optimization problem with box-constraints:

$$\text{find } \mathbf{x}^{**} \in \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmin}}\, f(\mathbf{x}) \tag{1}$$

where the search domain $\mathcal{X} = [-1, 1]^D$ is possibly of very high-dimensionality, say up to billions of variables. In all the rest, we suppose that $f^{**} = \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ exists.

Recently, a groundbreaking approach was proposed in the paper [55][1], that relies on random embeddings. Under the hypothesis that the efficient dimension of the problem is much smaller than the number of inputs, high-dimensional optimization problems can be provably solved using global optimization algorithms in moderate dimensions, relying on random embeddings and related results from random matrix theory. However implementing such algorithms in the case of bounded domains can be quite inefficient if the low-dimensional domain is not carefully chosen in accordance with the embedding, and if space deformations caused by such embedding are not accounted for. Here we tackle both issues, with a main focus on the choice of the low-dimensional domain, and instantiation in the BO framework. Before developing the main results and applications in Sections 2 and 3, let us present selected state-of-the-art works, and narrow down the present aim and scope in Sections 1.1 and 1.2 respectively.

1.1 Related works and the random embedding approach

Many authors have focused on methods to handle high-dimensionality, in several directions. Selecting few variables is a rather natural idea to get back to a moderate search space, as done e.g., in [48], [41], [44] or [5]. Another common strategy of dimension reduction is to construct a mapping from the high-dimensional research space to a smaller one, see e.g., [52], [33] and references therein. Other techniques suppose that the black-box function is only varying along a low dimensional subspace, possibly not aligned with the canonical basis, such as in [10] or [18], using low rank matrix learning. With few unknown active variables, [3] proposes to combine compressed sensing with linear bandits for a more parcimonious optimization. Lastly, incorporating structural assumptions such as additivity within GP models

[1] see [54] for the extended journal version

is another angle of attack, see [13], [12], [14], [32], [53] and references therein. They enjoy a linear learning rate with respect to the problem dimensionality, as used e.g., in [27] or [25].

In most of the above references, a significant part of the budget is dedicated to uncover the structure of the black-box, which may impact optimization in the case of very scarce budgets. In [55], with the Random EMbedding Bayesian Optimization (REMBO) method, a radically different point of view was adopted, by simply relying on a randomly selected embedding. Even if the main hypothesis is again that the high-dimensional function only depends on a low-dimensional subspace of dimension d, the so-called *low effective dimensionality* property, no effort is dedicated to learn it. This strong property is backed by empirical evidence in many cases, see e.g., references in [55], [7] or in [24].

The principle is to embed a low dimensional set - usually a box - $\mathcal{Y} \subseteq \mathbb{R}^d$ to $\mathcal{X}$, using a randomly drawn matrix $\mathbf{A} \in \mathbb{R}^{D\times d}$ within the mapping ϕ: $\mathbf{y} \in \mathcal{Y} \to p_{\mathcal{X}}(\mathbf{A}\mathbf{y}) \in \mathcal{X}$, where $p_{\mathcal{X}}$ denotes the convex projection onto $\mathcal{X}$. By doing so, the initial search space $\mathcal{X}$ is reduced to a fraction (possibly the totality) of the embedded space $\mathcal{E} := \phi(\mathbb{R}^d)$, which remains fixed for the rest of the process. The corresponding transformed optimization problem writes:

$$\text{find } \mathbf{x}^* \in \operatorname*{argmin}_{\mathbf{x}\in\mathcal{E}} f(\mathbf{x}) \tag{2}$$

which may seem unpractical when formulated directly in terms of the $\mathcal{X}$ space but can be grasped more intuitively when parametrized in terms of the $\mathcal{Y}$ space:

$$(\mathcal{R}): \text{ find } \mathbf{y}^* \in \mathcal{Y} \subseteq \mathbb{R}^d \text{ such that } f(\phi(\mathbf{y}^*)) = f(\mathbf{x}^*).$$

Conditions such that $f^* := f(\mathbf{x}^*) = f^{**}$ are addressed in [55] and relaxed, e.g., in [43]. Notably, solutions coincide if the influential subspace is spanned by variables of $\mathcal{X}$, i.e., most variables do not have any influence. Here we rather focus on ensuring that there exists $\mathbf{y}^* \in \mathcal{Y}$ such that $f(\phi(\mathbf{y}^*)) = f^*$, through the definition of $\mathcal{Y}$ in problem $(\mathcal{R})$ as detailed in Section 1.2. Remarkably, any global optimization algorithm can potentially be used for solving $(\mathcal{R})$. In the application section we focus specifically on GP-based BO methods.

To fix ideas, Fig. 1 is an illustration of the random embedding principle as well as of the various sets mentioned so far. On this example with $D = 2$, the original function is defined on $\mathcal{X} = [-1, 1]^2$ with a single unknown active variable ($d = 1$). Even if the search for the optimum is restricted to the red broken line with problem (2), a solution to problem (1) can still be found under the settings of problem $(\mathcal{R})$.

1.2 Motivation: Limits of random embeddings

If the random embedding technique with problem $(\mathcal{R})$ has demonstrated its practical efficiency in several test cases [55, 2], it still suffers from practical difficulties, related to the definition of $\mathcal{Y}$. The first one, as discussed in [55], is non-injectivity due to the convex projection, i.e., distant points in $\mathcal{Y}$ may have the same image in $\mathcal{E}$. In Fig. 1, this is the case with $\mathcal{Y}_1$, for the portions between the diamonds and crossed boxes. As a consequence, taking $\mathcal{Y}$ too large makes the search of solutions harder. Preventing this issue is possible by several means, such as using high-dimensional distance information. For GPs, it amounts to consider distances either on $\mathcal{E}$ within the covariance kernel as originally suggested, or, as proposed by

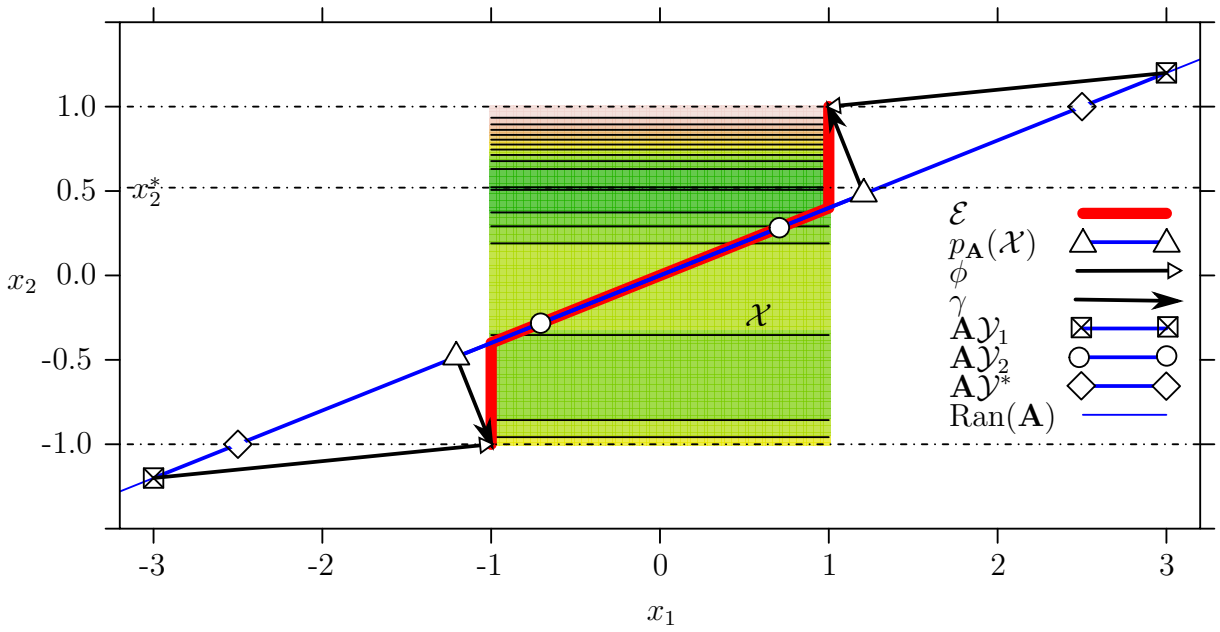


Fig. 1: Example with $d = 1$ and $D = 2$. The filled level lines are those of a function defined on $\mathcal{X}$ depending only on its second variable, whose optimal value is highlighted by the dashed line (at $x_2 \approx 0.52$). The image of the sets $\mathcal{Y}_1 = [-6, 6]$, $\mathcal{Y}_2 = [-1, 1]$ and $\mathcal{Y}^* = [-5, 5]$ by the matrix $\mathbf{A} = [0.5, 0.2]^\top$ are delimited on $\mathrm{Ran}(\mathbf{A})$ with symbols. The extent of the orthogonal projection of $\mathcal{X}$ onto $\mathrm{Ran}(\mathbf{A})$ is delimited by triangles.

[2], on the orthogonal projection $p_{\mathbf{A}}$ of $\mathcal{E}$ onto the random linear subspace spanned by $\mathbf{A}$, i.e., onto $\mathrm{Ran}(\mathbf{A})$. The advantage of the latter is to remain low-dimensional.

The example in Fig. 1 highlights another remaining difficulty, also mentioned e.g., in [27, 47]: if $\mathcal{Y}$ is too small (e.g., with $\mathcal{Y}_2$), there may be no solution to $(\mathcal{R})$. As of now, only empirical rules have been provided, with fixed bounds for $\mathcal{Y}$ (giving the too small $\mathcal{Y}_2$ in this example). A possible workaround is to use several embeddings, preferably in parallel [55]. We argue that it may also be a sound option to keep a single one, benefiting of parallelism in solving problem (2) instead.

1.3 Contributions

For a given random matrix $\mathbf{A}$, the selection of a set $\mathcal{Y}$ in solving problem $(\mathcal{R})$ balances between efficiency (in favor of a smaller search space), practicality (in its description) and robustness (i.e., contains a solution of (2)). Consequently, we focus on the question $(\mathcal{Q})$ of taking $\mathcal{Y}$ as a set of smallest volume still ensuring that solutions of $(\mathcal{R})$ and (2) are equivalent, i.e.,

$$(\mathcal{Q}) : \text{find } \mathcal{Y}^* \text{ such that } \mathrm{Vol}(\mathcal{Y}^*) = \inf_{\mathcal{Y} \subset \mathbb{R}^d,\ \phi(\mathcal{Y}) = \mathcal{E}} \mathrm{Vol}(\mathcal{Y}).$$

Based on an extensive description of sets related to the mapping, we exhibit a solution of $(\mathcal{Q})$, the set $\mathcal{U}$, while providing additional insight on the difficulties encountered in practice. In Fig. 1, the unique (in this $d = 1$ case) optimal $\mathcal{Y}^* = \mathcal{U} = [-5, 5]$ maps to the portion of $\mathrm{Ran}(\mathbf{A})$ between the two diamonds, whose convex projections on $\mathcal{X}$ are the extremities of $\mathcal{E}$

(as is also the case for any point further away from O). Unfortunately this strategy does not generalize well in higher dimensions, where the description of $\mathcal{U}$ becomes cumbersome, due to the convex projection component of the mapping ϕ.

A first attempt to alleviate the impact of the convex projection step proposed by [2] is to rely on an additional orthogonal projection step. Here, we extend this approach to completely by-pass the convex projection. That is, instead of associating a point of $\mathrm{Ran}(\mathbf{A})$ with its convex projection on $\mathcal{X}$, we propose to associate a point of $p_{\mathbf{A}}(\mathcal{E}) \subset \mathrm{Ran}(\mathbf{A})$ with its pre-image in $\mathcal{E}$ by $p_{\mathbf{A}}$, informally speaking *inverting* the orthogonal projection – *back-projecting* in short. Expressed in a basis of $\mathrm{Ran}(\mathbf{A})$, denoted by the $d \times D$ matrix $\mathbf{B}$, coordinates are d-dimensional. Then the embedding procedure is still a mapping between $\mathbb{R}^d$ and $\mathcal{E}$, denoted by $\gamma : \mathbf{B}p_{\mathbf{A}}(\mathcal{E}) \subset \mathbb{R}^d \to \mathcal{E}$. Most of the present work is dedicated to study the validity and applicability of this back-projection. The corresponding alternative formulations of $(\mathcal{R})$ and $(\mathcal{Q})$ write:

$$(\mathcal{R}') : \text{ find } \mathbf{y}^* \in \mathcal{Y} \subseteq \mathbb{R}^d \text{ such that } f(\gamma(\mathbf{y}^*)) = f^*$$

and

$$(\mathcal{Q}') : \text{find } \mathcal{Y}^* \text{ such that } \mathrm{Vol}(\mathcal{Y}^*) = \inf_{\mathcal{Y} \subset \mathbb{R}^d,\ \gamma(\mathcal{Y}) = \mathcal{E}} \mathrm{Vol}(\mathcal{Y}).$$

The benefits of these formulations are, first, that a solution of $(\mathcal{Q}')$ is by construction $\mathcal{Y}^* = \mathcal{Z} := \mathbf{B}p_{\mathbf{A}}(\mathcal{X})$ and, second, that they enjoy better properties from an optimization perspective. Back to Fig. 1, $p_{\mathbf{A}}(\mathcal{X})$ is delimited by the two triangles.

Finally, these enhancements are adapted for Bayesian optimization, via the definition of appropriate covariance kernels. They achieve significant improvements, with comparable or better performance than the initial version of REMBO on a set of examples, while discarding the risk of missing the optimum in $\mathcal{E}$.

The remainder of the article is as follows. In the subsequent Section 2 are presented our main results towards question $(\mathcal{Q})$, both in its original setup and in the alternative one $(\mathcal{Q}')$. A main contribution of this work is to explicitly write the sets $\mathcal{U}$, $\mathcal{Z}$, and the alternative mapping γ, depending on the matrix $\mathbf{A}$. While these results are of general interest for global optimization with random embedding regardless of the base optimization algorithm, Section 3 is dedicated to the particular case of Bayesian optimization with random embeddings on various experiments. Section 4 concludes the article.

2 Minimal sets solutions to questions $(\mathcal{Q})$ and $(\mathcal{Q}')$

Throughout this section, we consider that the matrix $\mathbf{A}$ is given, and for simplicity that it belongs to the following class of matrices, denoted $\mathcal{A}$:

$$\mathcal{A} = \left\{ \mathbf{A} \in \mathbb{R}^{D \times d} \text{ such that any } d \times d \text{ extracted submatrix is invertible (i.e., of rank } d) \right\}.$$

This mild condition is ensured with probability one for random matrices with standard Gaussian i.i.d. entries, as used in [55]. Before discussing the relative merits of problems $(\mathcal{R})$ and $(\mathcal{R}')$, we start by exhibiting sets of interest in $\mathcal{Y}$, $\mathcal{X}$ and $\mathrm{Ran}(\mathbf{A})$ in both cases.

2.1 A minimal set in $\mathbb{R}^d$ mapping to $\mathcal{E}$

Until now, the description of the set $\mathcal{Y}$ has been relatively vague – [54] states that there is room for improvement. This question is settled in [55] by setting $\mathcal{Y} = [-\sqrt{d}, \sqrt{d}]^d$, while [55, Theorem 3] only ensures to find a solution to problem $(\mathcal{R})$ with probability 1 in the particular case where $\mathcal{Y} = \mathbb{R}^d$. On the other hand, the sets $\mathcal{E}$ and $p_{\mathbf{A}}(\mathcal{X})$ are fixed and well defined given $\mathbf{A}$. This motivates us to describe a new set $\mathcal{U} \subset \mathbb{R}^d$, containing a solution of $(\mathcal{R})$, of minimal volume, and that can also be described from $\mathbf{A}$.

To this end, consider the low-dimensional space $\mathbb{R}^d$. Denote by $H_{\mathbf{a},\delta}$ the hyperplane in $\mathbb{R}^d$ with normal vector $\mathbf{a} \in \mathbb{R}^d$ and offset $\delta \in \mathbb{R}$: $H_{\mathbf{a},\delta} = \left\{\mathbf{y} \in \mathbb{R}^d \,,\, \langle \mathbf{a}, \mathbf{y} \rangle = \delta\right\}$. Our analysis in the low dimensional space begins by a general definition of *strips*.

Definition 2.1. *We call* strip *with parameters $a \in \mathbb{R}^d$ and $\delta \in \mathbb{R}$, denoted by $\mathcal{S}_{\boldsymbol{a},\delta}$ the set of points between the parallel hyperplanes $H_{\boldsymbol{a},-\delta}$ and $H_{\boldsymbol{a},\delta}$: $\mathcal{S}_{\boldsymbol{a},\delta} = \left\{\mathbf{y} \in \mathbb{R}^d \,,\, |\langle \boldsymbol{a}, \mathbf{y} \rangle| \leq |\delta|\right\}$.*

Let us now consider hyperplanes with normal vectors given by the rows of a matrix $\mathbf{A} \in \mathbb{R}^{D \times d}$ and with fixed $\delta = 1$. The D corresponding strips, now simply denoted $\mathcal{S}_i$, are given by:

$$\mathcal{S}_i = \left\{\mathbf{y} \in \mathbb{R}^d, -1 \leq \mathbf{A}_i \mathbf{y} \leq 1\right\}.$$

The intersection of all strips $\mathcal{S}_i$ is denoted $\mathcal{I}$. It corresponds to the pre-image of $\mathcal{X} \cap \mathrm{Ran}(\mathbf{A})$ by $\mathbf{A}$:

$$\mathcal{I} = \bigcap_{i=1}^{D} \mathcal{S}_i = \left\{\mathbf{y} \in \mathbb{R}^d, \forall i = 1, \ldots, D : -1 \leq \mathbf{A}_i \mathbf{y} \leq 1\right\} = \left\{\mathbf{y} \in \mathbb{R}^d \,,\, p_{\mathcal{X}}(\mathbf{A}\mathbf{y}) = \mathbf{A}\mathbf{y}\right\}.$$

Of interest will also be intersections of d strips, corresponding to *parallelotopes*, i.e., linear transformations of a d-cube in a d-dimensional subspace, see e.g., [31]. In particular, using set notations $I = \{i_1, \ldots, i_d\} \subseteq \{1, \ldots, D\}$, denote the parallelotope with strips I

$$\mathcal{P}_I = \left\{\mathbf{y} \in \mathbb{R}^d, \forall i \in I : -1 \leq \mathbf{A}_i \mathbf{y} \leq 1\right\} = \bigcap_{i \in I} \mathcal{S}_i.$$

There are $\binom{D}{d}$ different parallelotopes $\mathcal{P}_I$. We thus consider their union, which is referred to as $\mathcal{U}$:

$$\mathcal{U} = \bigcup_{I \subseteq \{1,\ldots,D\}, |I| = d} \mathcal{P}_I$$

where $|I|$ is the size of I. In fact we show in the following Theorem 1 that $\mathcal{U}$ is the smallest closed set such that the map $\phi_{|\mathcal{U}} : \mathcal{U} \to \mathcal{E}$, $\mathbf{y} \mapsto p_{\mathcal{X}}(\mathbf{A}\mathbf{y})$ is surjective.

Theorem 1. *If $\mathbf{A} \in \mathcal{A}$, then $\mathcal{U}$ is the smallest closed set $\mathcal{Y} \subseteq \mathbb{R}^d$ such that $p_{\mathcal{X}}(\mathbf{A}\mathcal{Y}) = \mathcal{E}$. Furthermore, $\mathcal{U}$ is a compact and star-shaped set with respect to every point in $\mathcal{I}$.*

The detailed proof is can be found in Appendix A.2.

In other words, if we choose $\mathcal{Y} \supseteq \mathcal{U}$, then a solution to problem $(\mathcal{R})$ and (2) can be found. Yet, from its definition as a union of intersections, the set $\mathcal{U}$ is unpractical to directly work with. Indeed, it is more common practice to work on simpler sets such as boxes instead of

star-shaped sets. Based upon [55, Theorem 3], their choice of $\mathcal{Y} = [-\sqrt{d}, \sqrt{d}]^d$ originates from results on the radius of a parallelotope $\mathcal{P}_I$, corresponding to a probability greater than $1-\varepsilon$ of containing a solution, with $\varepsilon = \log(d)/\sqrt{d}$ (see the discussion in [54]). Unfortunately, there is no such result for any matrix in general nor for the maximum radius of a parallelotope in which one could wish to enclose $\mathcal{U}$. Yet, it is still easy to detect whether a given point is in $\mathcal{U}$ or not: if no more than d components of $\mathbf{Ay}$ are superior to one in absolute value. Hence selecting a large $\mathcal{Y}$ such that $\mathcal{U} \subseteq \mathcal{Y}$ is always possible, even though it may prove to be extremely large and counterproductive.

As a by-product, the proof of Theorem 1 gives a possibility to find pre-images in $\mathcal{Y}$ for elements of $\mathcal{E}$: letting $\mathbf{x} \in \mathcal{E}$, pre-image(s) in $\mathcal{U}$ are solutions of the following system of linear (in)equations:

$$\text{find } \mathbf{y} \in \mathcal{U} \text{ s.t. } \begin{cases} \mathbf{A}_J\mathbf{y} & = \mathbf{x}_J \\ \mathbf{A}_K\mathbf{y} & \geq \mathbf{1}_{|K|} \\ \mathbf{A}_L\mathbf{y} & \leq -\mathbf{1}_{|L|} \end{cases}$$

where J, K and L are the sets of components of $\mathbf{x}$ such that $|x_i| \leq 1$, $x_i > 1$ and $x_i < -1$ respectively, with $\mathbf{1}_{|K|} = (1, \ldots, 1)^\top$ of length $|K|$. A solution to this problem exists by Theorem 1, several may even exist if $|J| < d$.

To sum up, we have highlighted three different sets of interest: parallelotopes $\mathcal{P}_I$, the intersection of all of them $\mathcal{I}$, and their union $\mathcal{U}$. The sets $\mathcal{U}$, $\mathcal{P}_I$ and $\mathcal{I}$ are illustrated with $d = 2$ in Fig. 2. On the top figures, (a), strips are marked by lines. Next, we conduct a similar analysis of with the mapping γ.

2.2 Bijection between $\mathcal{E}$ and $\mathcal{Z}$

The core idea behind formulations $(\mathcal{R}')$ and $(\mathcal{Q}')$ is, through using $\text{Ran}(\mathbf{A})$ as low-dimensional domain, to replace the convex projection by an inversion of the orthogonal projection, *in fine* replacing the mapping ϕ by another, γ, with better properties. It is worth insisting that the search for a minimum occurs on the same set $\mathcal{E}$ in the original high-dimensional domain $\mathcal{X}$.

A core set here is the one obtained by projection of $\mathcal{X}$ onto $\text{Ran}(\mathbf{A})$, which is known to be a zonotope, a special class of convex centrally symmetric polytopes, see Definition 2.2 and e.g., in [36], [56] or [31].

Definition 2.2 (Zonotope as hypercube projection, adapted from [31])**.** *A (centered) D-zonotope in $\mathbb{R}^d$ is the image of the $[-1, 1]^D$ hypercube $\mathcal{X}$ by a linear mapping. Given a matrix $\mathbf{B} \in \mathbb{R}^{d \times D}$ representing the linear mapping, the zonotope $\mathcal{Z}$ is defined by $\mathcal{Z} = \mathbf{B}\mathcal{X}$.*

This representation of a zonotope is known as its *generator* representation, while it can also be described by vertices enumeration or hyperplane intersections like any other convex set, see e.g., [28]. They provide a very compact representation of sets, which is useful for instance in set estimation, see e.g., [31].

In the following, we assume that rows of $\mathbf{B}$ form an orthonormal basis of $\text{Ran}(\mathbf{A})$ in $\mathbb{R}^D$, i.e., $\mathbf{B}\mathbf{B}^\top = \mathbf{I}_d$, with $\mathbf{I}_d$ the identity matrix of size $d \times d$. The orthogonal projection onto $\text{Ran}(\mathbf{A})$ then simply writes: $p_{\mathbf{A}} = \mathbf{B}^\top\mathbf{B}$ [37]. Let us point out that without the

orthonormality condition, expressions involve pseudo-inverses. Now, as we aim to define a mapping between $\mathcal{Z}$ and $\mathcal{E}$, a key element given by Proposition 2.1 is that the orthogonal projection of the set $\mathcal{E}$ onto $\mathrm{Ran}(\mathbf{A})$ actually coincides with the one of $\mathcal{X}$ onto $\mathrm{Ran}(\mathbf{A})$, i.e., $\mathbf{B}^\top\mathcal{Z}$. It thus inherits the properties of a zonotope.

Proposition 2.1. *$p_{\mathbf{A}}(\mathcal{X}) = p_{\mathbf{A}}(\mathcal{E})$, or equivalently, $\mathbf{B}\mathcal{E} = \mathbf{B}\mathcal{X} = \mathcal{Z}$.*

Proof. Please refer to Appendix A.3. □

To provide some intuition about the ideas and sets involved in this Section 2.2, compared to those of Section 2.1, they are illustrated with an example in Fig. 2, panel (d).

Now, the difficulty for the associated mapping is to *invert* the orthogonal projection of $\mathcal{E}$ onto $\mathrm{Ran}(\mathbf{A})$, more precisely onto an orthonormal basis $\mathbf{B}$ of the latter. One way to perform this task is to define $\gamma(\mathbf{y}) : \mathcal{Z} \to \mathbb{R}^D$ as the map that, first linearly embeds $\mathbf{y} \in \mathcal{Z}$ to $\mathbb{R}^D$ with the matrix $\mathbf{B}^\top$ and then maps this $\mathbf{B}^\top\mathbf{y} \in \mathbb{R}^D$ to the closest point $\mathbf{x} \in \mathcal{X}$ whose orthogonal projections onto $\mathrm{Ran}(\mathbf{A})$, i.e., $\mathbf{Bx}$, coincide with $\mathbf{B}^\top\mathbf{y}$. This is represented in Fig. 1 with arrows, in the illustrative case $d = 1$, $D = 2$.

Now let us show that the map γ is well defined. Let $p_{\mathbf{A}}^{-1}(\mathbf{a}) = \left\{\mathbf{x} \in \mathbb{R}^D \text{ s.t. } p_{\mathbf{A}}(\mathbf{x}) = \mathbf{a}\right\}$ the set of pre-images of $\mathbf{a} \in \mathbb{R}^D$ for the orthogonal projection onto $\mathrm{Ran}(\mathbf{A})$. Then,

$$\gamma(\mathbf{y}) = p_{\mathcal{X} \cap p_{\mathbf{A}}^{-1}(\mathbf{B}^\top\mathbf{y})}(\mathbf{B}^\top\mathbf{y})$$

is the convex projection on the convex set $\mathcal{X} \cap p_{\mathbf{A}}^{-1}(\mathbf{B}^\top\mathbf{y})$. Since $\mathbf{y} \in \mathcal{Z}$, $\mathcal{X} \cap p_{\mathbf{A}}^{-1}(\mathbf{B}^\top\mathbf{y}) \neq \emptyset$ and γ is defined. This leads to the counterpart of Theorem 1 with the mapping γ.

Theorem 2. *$\mathcal{Z}$ is the smallest closed set $\mathcal{Y} \subseteq \mathbb{R}^d$ such that $\gamma(\mathcal{Y}) = \mathcal{E}$. Furthermore, $\mathcal{Z}$ is a compact, convex and centrally symmetric set.*

Proof. Please refer to Appendix A.4. □

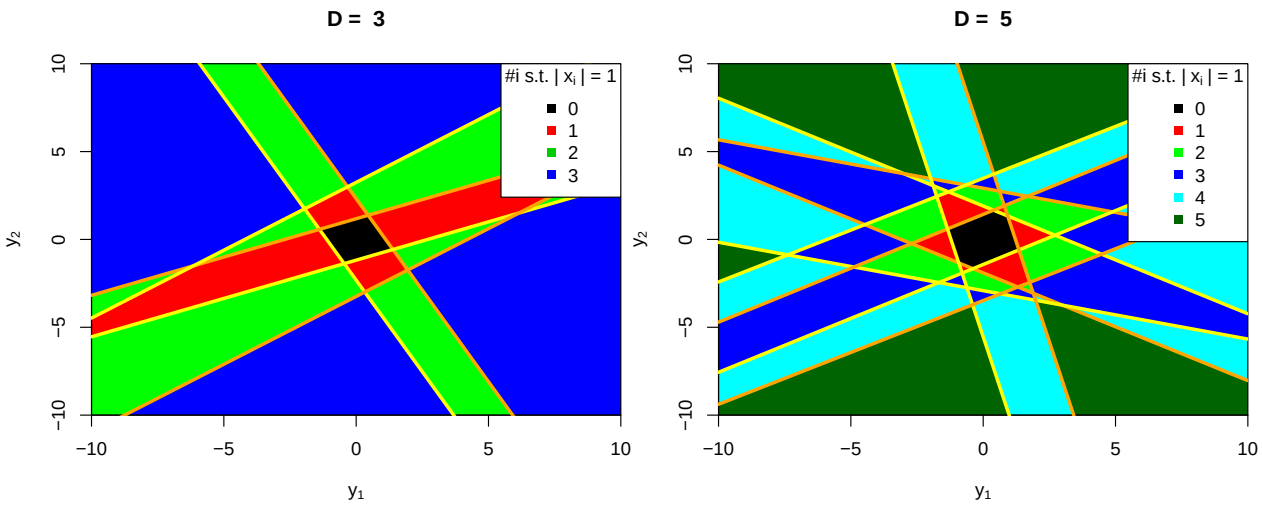
In practice, γ can be written as the solution of the following quadratic programming problem:

$$\gamma(\mathbf{y}) = \arg\min_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - \mathbf{B}^\top\mathbf{y}\|^2 \quad s.t.\ \mathbf{Bx} = \mathbf{y}.$$

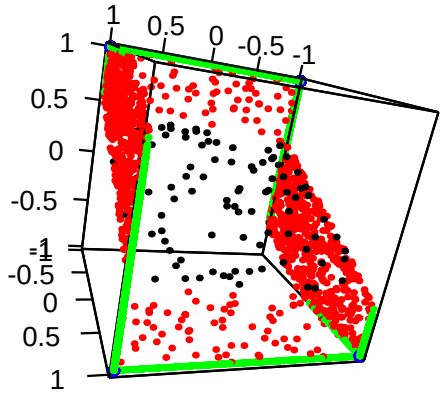
Next we discuss the relative merits of both solutions sets $\mathcal{U}$ and $\mathcal{Z}$, with mappings ϕ and γ respectively.

2.3 Discussion

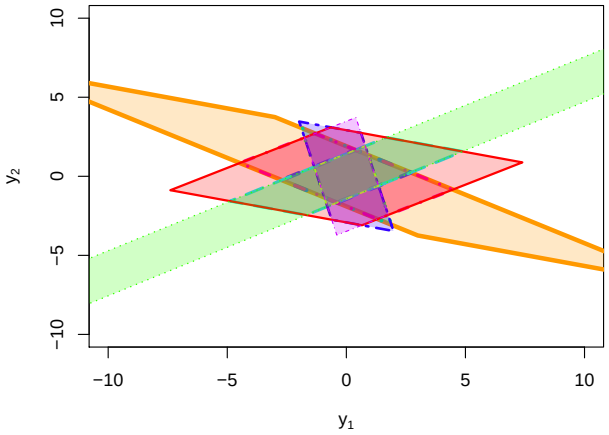
Recall that there is a compromise between practical implementation, size of the domain (related to convergence speed) and risk of missing a solution. The original REMBO, with ϕ and $\mathcal{Y} = [-\sqrt{d}, \sqrt{d}]^d$ is computationally efficient with its mapping procedure and a fixed definition of the search space. The balance between the two other points depends on the sampled matrix $\mathbf{A}$, and is expected to favor small domains [55].



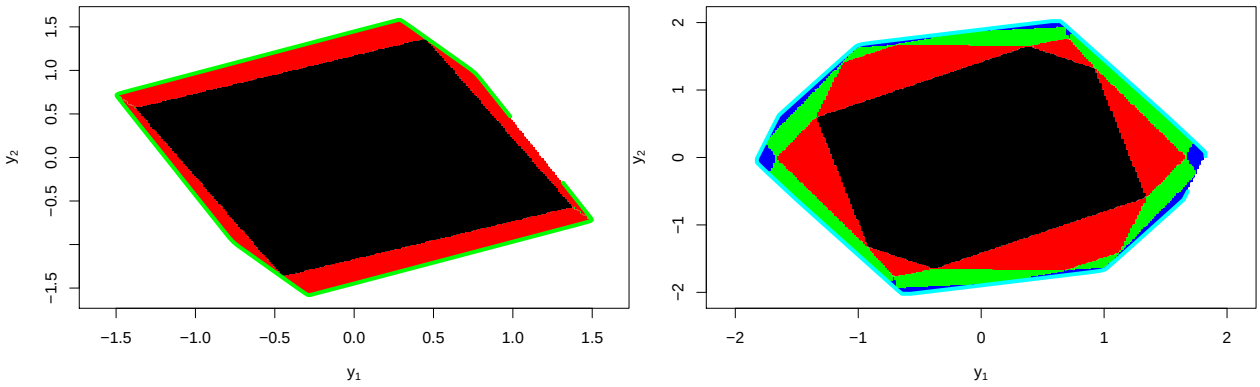
(a) Strips $\mathcal{S}_i$ in $\mathbb{R}^2$, formed by the orange and yellow lines standing for the parallel hyperplanes.



(b) $\mathcal{E}$ in $\mathbb{R}^3$.



(c) Parallelotopes in $\mathbb{R}^2$ forming $\mathcal{U}$.



(d) Zonotopes $\mathcal{Z}$ in $\mathbb{R}^2$.

Fig. 2: Representation of the sets of interest introduced in Section 2 with $d = 2$ and $D = 3$ (left) or $D = 5$ (right) for fixed $\mathbf{A}$ matrices. Except for panel (c), colors represents how many variables in $\mathcal{X}$ are not in $]-1, 1[$. The intersection of all parallelograms $\mathcal{I}$ is in black. Panel (a) highlights strips in $\mathcal{Y}$ when using the mapping ϕ. Panel (b) is the image by ϕ of the top left panel, in $\mathcal{X}$. Panel (c): each of the 10 parallelotopes (here parallelograms) $\mathcal{P}_I$ is depicted with a different color and their union is the minimal set $\mathcal{U}$. For illustration purpose, $\mathcal{U}$ is truncated: the cut green parallelogram is approximately enclosed in $[-210, 210] \times [130, 130]$. Panel (d): zonotopes used as pre-images with mapping γ.

If, as with questions $(\mathcal{Q})$ and $(\mathcal{Q}')$, emphasis is on ensuring that there is a solution in the low dimensional search space, and even if both sets $\mathcal{U}$ and $\mathcal{Z}$ are compacts, $\mathcal{Z}$ has several advantages. First, it is a convex set instead of a star-shaped one, with a generator description instead of a combinatorially demanding union description. Enclosing $\mathcal{U}$ in a box or a sphere requires finding the radius of the largest parallelope that enclose it, which is combinatorially difficult. Of interest here, the smallest box enclosing $\mathcal{Z}$ has a simple expression: the extreme value in the i^{th} direction is $\sum_{j=1}^{D} |B_{i,j}|$, $1 \leq i \leq d$, see e.g., in [31]. Hence it is possible to work in

$$\mathcal{B} = \left[-\sum_{j=1}^{D} |B_{1,j}|, \sum_{j=1}^{D} |B_{1,j}| \right] \times \cdots \times \left[-\sum_{j=1}^{D} |B_{d,j}|, \sum_{j=1}^{D} |B_{d,j}| \right].$$

Testing whether or not a given point $\mathbf{y} \in \mathbb{R}^d$ is in $\mathcal{Z}$ amounts to verify whether or not the linear system $\mathbf{B}\mathbf{x} = \mathbf{y}$ has a solution in $\mathcal{X}$; more conditions such as to identify the boundary of $\mathcal{Z}$ are given e.g., in [4]. As for $\mathcal{U}$, it amounts to verify that at least d variables of $\mathbf{A}\mathbf{y}$ are in $[-1, 1]$.

Even if γ requires solving a quadratic programming problem, the additional cost usually fades in the case of expensive black-box simulators, with limited budget of evaluations. Finally, additional advantages of $\mathcal{Z}$ over $\mathcal{U}$ in practice are provided in Proposition 2.2.

Proposition 2.2. *Denote* Vol_d *the* d*-volume in* $\mathbb{R}^D$*. Let* $\mathbf{A} \in \mathbb{R}^{D \times d}$ *with orthonormal columns, i.e.,* $\mathbf{A}^\top \mathbf{A} = \mathbf{I}_d$*. Then* $\mathrm{Vol}_d(\mathbf{A}\mathcal{U}) \geq \mathrm{Vol}_d(\mathcal{E}) \geq \mathrm{Vol}_d(\mathbf{A}\mathcal{Z})$*. If, in addition, rows of* $\mathbf{A}$ *have equal norm, then* $\mathrm{Vol}(\mathcal{Z}) / \mathrm{Vol}(\mathcal{I}) \leq d^{d/2}$*, which does not depend on* D*.*

Proof. Please refer to Appendix A.5. □

This result provides us with hints on the deformation of the search space occurring with both mappings, sharing the same invariant space $\mathcal{I}$. With ϕ, the volume of $\mathbf{A}\mathcal{U} \setminus \mathbf{A}\mathcal{I}$ is relatively bigger than the one of $\mathcal{E} \setminus \mathbf{A}\mathcal{U}$. The effect is reversed for γ. Since $\mathcal{I}$ is the set where all variables, especially the relevant ones, are not fixed[2], focusing more on these is, arguably, beneficial. In fact, the second part of the result indicates that the relative volume of the undeformed part within $\mathcal{Z}$ does not depend on D. In preliminary tests, we tried to extend the domain to $\mathcal{U}$ with the mapping ϕ, and the performance degraded considerably. In particular, the volume of $\mathcal{I}$ became quickly negligible compared to the one of $\mathcal{U}$, when increasing d or D. We next discuss these points in more details for GP-based BO, illustrating these differences empirically.

Remark 2.1. *In [55],* $\mathbf{A}$ *is a random matrix with i.i.d. standard Gaussian entries. In this case, many results about determinants, eigenvalues and limiting distributions are known, see e.g., [50] or [42] as starting points into this rich literature. One result related to Proposition 2.2 is that* $\|\frac{1}{D}\mathbf{A}^\top \mathbf{A} - \mathbf{I}_d\| \to 0$ *almost surely as* d/D *goes to 0, see e.g., [50]. There are several alternatives to Gaussian random matrices, such as random matrices whose rows are randomly selected on the sphere – corresponding to random matrices with independent rows of equal norm – that have the same asymptotic properties [50]. Their use has been studied in [1], showing benefits mostly for small* d*.*

[2]Outside of the set corresponding to $\mathcal{P}_I$ in $\mathbb{R}^d$, these influential variables I would be fixed to ± 1.

3 Application to Bayesian optimization

The random embedding paradigm incorporates seamlessly within GP-based BO methods through the covariance kernel. After a brief description of Bayesian optimization using Gaussian processes and the specific choices of covariance kernels, we present results on a set of test cases.

3.1 Modified REMBO procedure

Bayesian optimization, and especially seminal works on the expected improvement initiated in [39], is built on two key concepts: the first one is to consider the underlying black-box function as random and to put a prior distribution that reflects beliefs about it. New observations are used to update the prior, resulting in a posterior distribution. The second pillar is an acquisition function that selects new locations using the posterior distribution to balance exploitation of promising areas and exploration of less known regions.

One such example is the widely used EGO algorithm [26]. Its prior distribution is a Gaussian process, and its acquisition function the Expected Improvement (EI) criterion. Other popular surrogate models include radial basis functions, random forests and neural networks, see e.g., [19, 23, 29, 6] and references therein. As for alternative acquisition functions, we also mention those relying on an information gain as in [51, 21]. The reader interested in these variations on BO may refer to [47] for a recent review.

GP priors are attractive for their tractability since they depend only on a mean $\mu(\cdot)$ and covariance function $k(\cdot,\cdot)$. Assuming that μ and k are given, conditionally on n observations of f, $\mathbf{f}_{1:n} := (f(\mathbf{x}_1),\dots,f(\mathbf{x}_n))$, the predictive distribution is another GP, with mean and covariance functions given by:

$$m_n(\mathbf{x}) = \mu(\mathbf{x}) + \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1}(\mathbf{f}_{1:n} - \mu(\mathbf{x}_{1:n})) \tag{3}$$

$$k_n(\mathbf{x},\mathbf{x}') = k(\mathbf{x},\mathbf{x}') - \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1}\mathbf{k}(\mathbf{x}') \tag{4}$$

where $\mathbf{x}_{1:n} := (\mathbf{x}_1,\dots,\mathbf{x}_n)^\top$, $\mathbf{k}(\mathbf{x}) := (k(\mathbf{x},\mathbf{x}_1),\dots,k(\mathbf{x},\mathbf{x}_n))^\top$ and $\mathbf{K} := (k(\mathbf{x}_i,\mathbf{x}_j))_{1\le i,j\le n}$ are the vector of covariances of $Y(\mathbf{x})$ with the $Y(\mathbf{x}_i)$'s and the covariance matrix of $Y(\mathbf{x}_{1:n})$, respectively. The choice of the mean and covariance function dictates the expected behavior of f. Commonly, μ is supposed to be constant while k belongs to a parametric family of covariance functions such as the Gaussian and Matérn kernels, corresponding to different hypothesis about the derivability of f. Associated hyperparameters are frequently inferred based on maximum likelihood estimates, see e.g., [44] or [46] for specific details.

In the case of EI, the improvement is defined as the difference between the current minimum of the observations and the new function value (thresholded to zero). The latter being given by the GP model, EI is the conditional expectation of the improvement at a new observation, which has a closed form expression, see e.g., [26]. Notice that optimizing EI may be a complicated task in itself, see e.g., [17], due to multi-modality and plateaus. Yet evaluating EI is inexpensive and off-the-shelf optimization algorithms can be used (possibly relying on derivatives).

Adapting the framework of Bayesian optimization to incorporate a random embedding amounts to optimize the acquisition function on $\mathcal{Y}$, while evaluations are performed on $\mathcal{E}$. In terms of GP modeling, when using stationary covariance kernels, what matters is the distance between points. Several options are possible to account for high-dimensional distances through compositions of kernels with functions, also known as warpings. Existing warpings for k defined on $\mathcal{Y}$ include:

- identity warping: distances are distances in $\mathcal{Y}$, the corresponding kernel is denoted $k_{\mathcal{Y}}$ in [55];
- random embedding and convex projection warping, i.e., using ϕ, denoted $k_{\mathcal{X}}$ in [55];
- an additional composition is proposed by [2], with orthogonal projection onto $\mathrm{Ran}(\mathbf{A})$ and a distortion. The distortion is used to counteract the effect of the orthogonal projection on high dimensional distances: the further away from $\mathrm{Ran}(\mathbf{A})$, the closer to the center the projection is. The warping Ψ writes $\Psi(\mathbf{y}) = \left(1 + \frac{\|\phi(\mathbf{y}) - \mathbf{z}'\|}{\|\mathbf{z}'\|}\right)\mathbf{z}'$ with $\mathbf{z}' = \mathbf{z}/\max(1, \max_{1\leq i\leq D} |z_i|)$, $\mathbf{z} = p_{\mathbf{A}}(\phi(\mathbf{y}))$.

With the alternative mapping γ, $k_{\mathcal{Y}}$ is defined based on distances in $\mathcal{Z}$, while $k_{\mathcal{X}}$ makes use of γ instead of ϕ. As for k_Ψ, the orthogonal projection is already performed and it only amounts to applying the correction: $\Psi'(\mathbf{y}) = \left(1 + \frac{\|\gamma(\mathbf{y}) - \mathbf{z}'\|}{\|\mathbf{z}'\|}\right)\mathbf{z}'$ with $\mathbf{z}' = \mathbf{z}/\max(1, \max_{1\leq i\leq D} |z_i|)$, $\mathbf{z} = \mathbf{B}^\top\mathbf{y}$). For the sake of readability, consider an isotropic *squared-exponential* (or *Gaussian*) kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/\theta^2)$ where θ is the scale or "lengthscale" hyperparameter, Table 1 summarizes the expressions for all combinations of warping and mappings.

Table 1: Gaussian kernel expressions depending on the embedding and warping. The first column summarizes existing kernels in the literature [55, 2] relying on ϕ, the second their transposition when using γ.

	mapping ϕ $(\mathbf{y}, \mathbf{y}' \in \mathcal{Y})$	mapping γ $(\mathbf{y}, \mathbf{y}' \in \mathcal{Z})$
$\mathbb{R}^d$	$k_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}') = \exp(-\|\mathbf{y} - \mathbf{y}'\|^2/\theta^2)$	$k_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}') = \exp(-\|\mathbf{y} - \mathbf{y}'\|^2/\theta^2)$
$\mathcal{E}$	$k_{\mathcal{X}}(\mathbf{y}, \mathbf{y}') = \exp(-\|\phi(\mathbf{y}) - \phi(\mathbf{y}')\|^2/\theta^2)$	$k_{\mathcal{X}}(\mathbf{y}, \mathbf{y}') = \exp(-\|\gamma(\mathbf{y}) - \gamma(\mathbf{y}')\|^2/\theta^2)$
$\mathrm{Ran}(\mathbf{A})$	$k_\Psi(\mathbf{y}, \mathbf{y}') = \exp(-\|\Psi(\mathbf{y}) - \Psi(\mathbf{y}')\|^2/\theta^2)$	$k'_\Psi(\mathbf{y}, \mathbf{y}') = \exp(-\|\Psi'(\mathbf{y}) - \Psi'(\mathbf{y}')\|^2/\theta^2)$

To take into account that γ is not defined outside of $\mathcal{Z}$, since $\mathcal{B}$ is only employed to maximize EI as acquisition function, which is positive, we propose to define $EI_{\text{ext}} : \mathbb{R}^d \to \mathbb{R}$ using a penalization as follows:

$$EI_{\text{ext}}(\mathbf{y}) = \begin{cases} EI(\gamma(\mathbf{y})) \text{ if } \mathbf{y} \in \mathcal{Z} \\ -\|\mathbf{y}\| \text{ else} \end{cases}$$

where testing if $\mathbf{y} \in \mathcal{Z}$ is performed by checking if the linear system $\mathbf{Bx} = \mathbf{y}$ has a solution $\mathbf{x} \in \mathcal{X}$. The same test is to be used to build an initial design of experiments in $\mathcal{Z}$. The penalty $-\|\mathbf{y}\|$ if $\mathbf{y} \notin \mathcal{Z}$ has been chosen to push toward the center of domain, thus toward $\mathcal{Z}$. An outline of the resulting REMBO procedure with the proposed improvements is given in Algorithm 1.

Algorithm 1 Pseudo code of the REMBO procedure with mapping γ

Require: d, n_0, kernel k (e.g., among $k_{\mathcal{Y}}, k_{\mathcal{X}}, k_{\Psi}$).
Sample $\mathbf{A} \in \mathbb{R}^{D\times d}$ with independent standard Gaussian coefficients.
Apply Gram-Schmidt orthonormalization to $\mathbf{A}$.
Define $\mathbf{B} = \mathbf{A}^\top$ and compute $\mathcal{B}$.
Construct an initial design of experiment in $\mathcal{Z}$, of size n_0.
Build the GP model with kernel k.
while time/evaluation budget not exhausted **do**
    Find $\mathbf{y}_{n+1} = \text{argmax}_{\mathbf{y}\in\mathcal{B}} EI_{\text{ext}}(\mathbf{y})$
    Evaluate the objective function at $\mathbf{y}_{n+1}$, $f_{n+1} = f(\gamma(\mathbf{y}_{n+1}))$.
    Update the GP model based on new data.
end while

3.2 Numerical experiments

We propose to illustrate the interest of the proposed modifications on several benchmark functions of various dimensionality, as summarized in Table 2. Some are classical multimodal synthetic functions such as Branin, Hartman6, Giunta and Levy, which are reasonably well modeled by Gaussian process. The Borehole function [40] models the water-flow in a borehole, commonly used in the computer experiments literature. The last one, the Cola function is a weighted least squares scaling problem, used for instance in multidimensional scaling, see, e.g., [34]. The number of influential dimensions of those problems varies from 2 to 17, for a total number of variables D ranging from 17 to 200. The former are chosen randomly, but kept fixed for each specific run (25 in total) to ensure fairness among comparators. The budget for optimization is either 100 or 250 evaluations, which are representative of expensive optimization tasks.

Table 2: Summary of test functions

name	d	d_e	D	name	d	d_e	D
Branin [9]	2	2	25, 100	Hartman6 [9]	6	6	50, 200
Giunta [38]	2	2	80	Borehole [40]	8	8	50
Cola [35]	6	17	17	Levy [30]	10	10	80

The emphasis is on both the average and worst case performances, as with the new formulation the search space $\mathcal{Z}$ maps with γ to the entire $\mathcal{E}$. We compare it to the original choice of REMBO: mapping ϕ with search domain $\mathcal{Y} = [-\sqrt{d}, \sqrt{d}]^d$. Preliminary tests with search domains encompassing estimated $\mathcal{U}$ showed a degraded performance compared to those with $\mathcal{Y}$ and are not reported here. We also test the three possible covariance kernels (see Table 1) with both mappings. Experiments have been performed relying on the `DiceKriging` and `DiceOptim` packages [46], with an unknown constant trend and Matérn covariance kernel with $\nu = 5/2$. The corresponding code is publicly available at https://github.com/mbinois/RRembo. For solving the quadratic programming problem within γ, we use the `quadprog` package [49]. In all the problems here, the corresponding extra cost was not more than a dozen of milliseconds per solve.

The baseline performance is given by uniform sampling (RO) in the original domain $\mathcal{X}$. We also compare to the Ensemble Bayesian Optimization method [53] using the `Python` code made available by the authors. It relies on an ensemble of additive GP models on randomized partitions of the space, which has been shown to scale both in terms of dimension and number of variables. We use the default tuning parameters, with batch size twenty, and let the number of additive components increase up to d.

We use the optimality gap as a metric, i.e., best value found minus known optimum value. The results are provided in Figs. 3 and 4, corresponding to final boxplots and progress over iterations, respectively. Overall, the median performance of REMBO variants is better than both uniform sampling (RO) and ensemble Bayesian optimization (EBO). Moreover, the worst performance in terms of 75% quantile is almost always improved with the mapping γ. Between the three kernel choices, k_Ψ is consistently a sound choice, while the performance of $k_\mathcal{X}$ is highly variable. As a result, looking at the best rank over all tests, γ with k_Ψ is the best combination.

The results are the most mitigated for the $d = 2$ cases, where the mapping ϕ can outperform the mapping γ. As d increases, the difference becomes more striking in favor of γ, and its 75% quartile is always below the 25% quartile from RO. In the Levy case, where $d = 10$, the original REMBO method is even worse than RO. Independently of the kernels, a proportion of under-performing outliers with mapping ϕ and fixed $\mathcal{Y} = [-\sqrt{d}, \sqrt{d}]^d$ can be interpreted as cases when the optimal solution is not contained in the domain; these do not happen with γ. For some of the remaining ones, the reason may be related to an unfavorable arrangements of strips for the GP modeling that could be alleviated with further work on kernels.

Figure 3 also shows differences in initial designs with respect to the mapping used. There is no clear trend since the best design strategy depends on the problem at hand and the location of the optima. For instance, in the case of Borehole, designs using ϕ on $\mathcal{Y} = [-\sqrt{d}, \sqrt{d}]^d$ are better starting points than with γ on $\mathcal{Z}$, but this advantage is quickly reduced. On a different aspect, shown on the Branin and Hartman6 functions, increasing D does not affect the performance of the REMBO methods more than the choice of the active dimensions. We did not conduct such a study for EBO. The Cola function illustrates the case when all D variables are influential. Even if this is not a favorable setup for REMBO, it still outperforms RO and EBO with limited budgets.

Under the limited budgets used here, relying on random embeddings proved much better than uniform sampling. The only exception – for the original method only – is with Levy, highlighting that the choice of the domain is crucial with respect to the performance of the method. It also illustrates that when the budget is low, it may be detrimental to balance observations to learn the structure of the function, such as with EBO. On all examples, considering problem $(\mathcal{R}')$ thus appears as a sound alternative to $(\mathcal{R})$. Indeed, initial concerns that a larger search space may impact the average performance do not reflect on the results, even often showing a superior performance. As for robustness, the worst performances have been greatly improved in general.

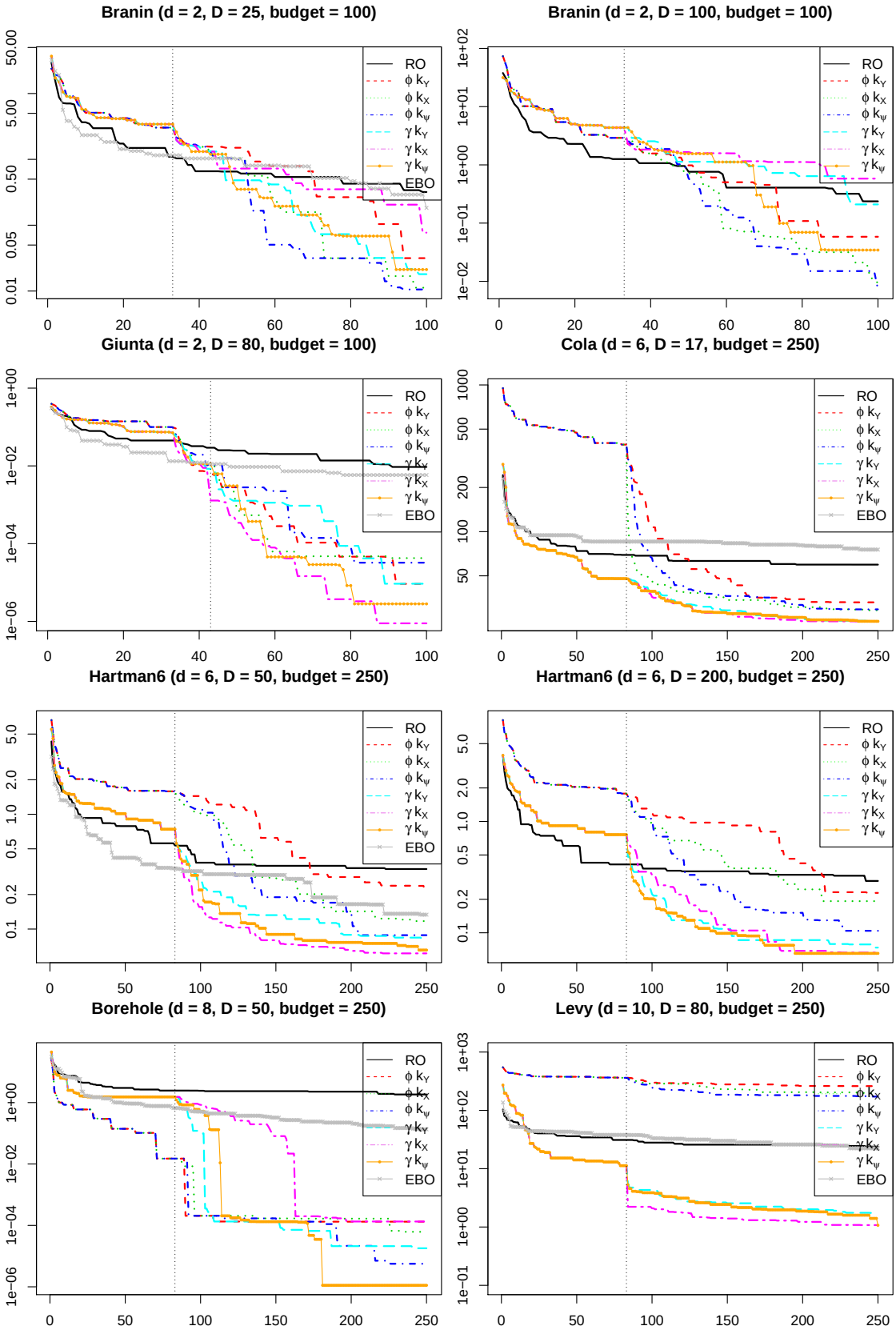


Fig. 3: Decrease of median optimality gap (log scale) over iterations for random optimization (RO), ensemble Bayesian optimization (EBO) as well as variants of REMBO with mappings ϕ (on $[-\sqrt{d}, \sqrt{d}]^d$) or γ (on $\mathcal{Z}$) and kernels in Table 1, on the test problems of Table 2. The dotted vertical lines marks the end of the design of experiments phase.

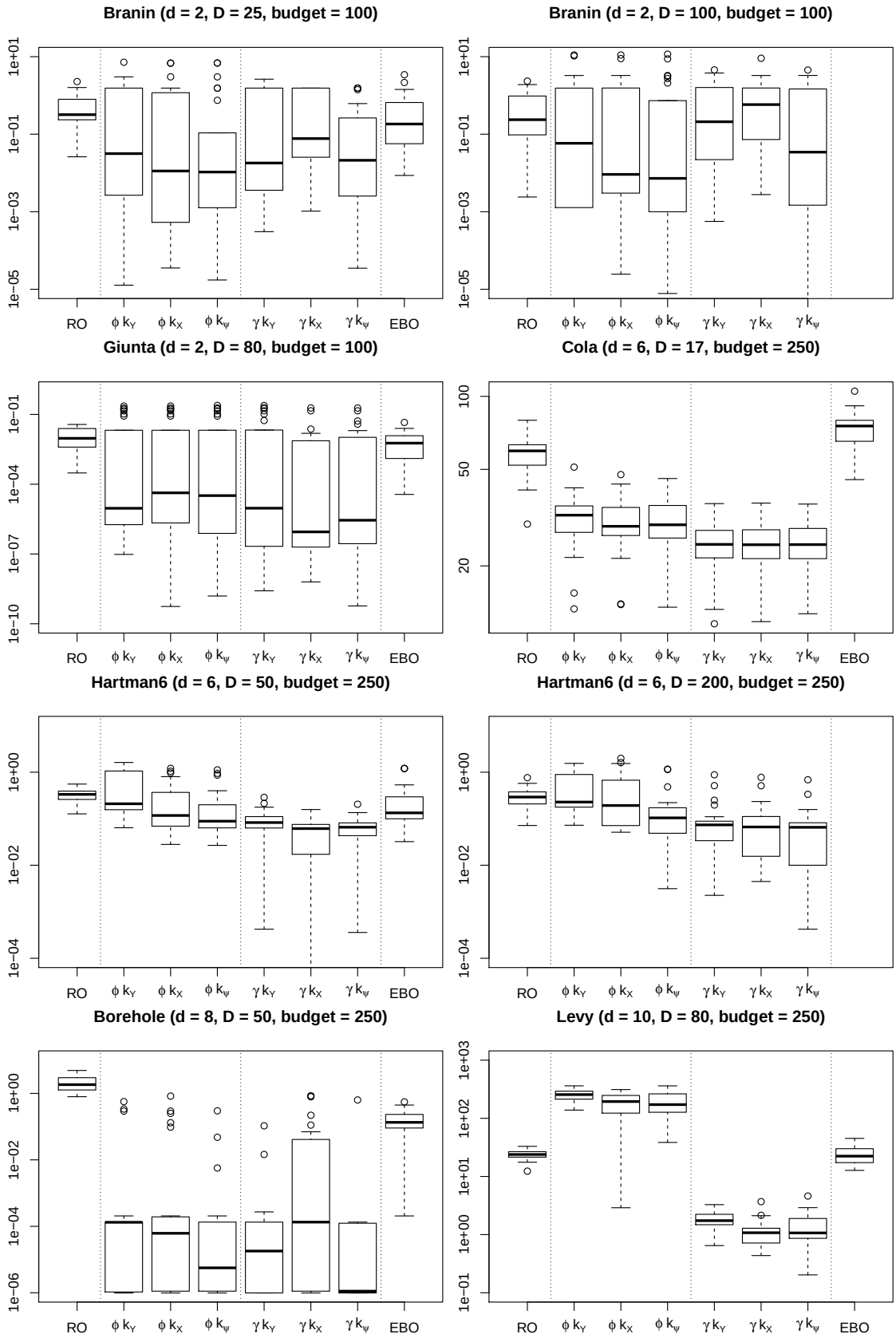


Fig. 4: Boxplots of optimality gap (log scale), corresponding to the last iteration in Figure 3, grouped by mappings.

4 Conclusion and perspectives

Although random embeddings offers a simple yet powerful framework to perform high-dimensional optimization, a snag relies in the definition of bounds for the low-dimensional search space. In the original setting, this results in an unsatisfactory compromise between hindering efficiency or taking the risk of discarding global solutions. While for this latter guarantees were given with probability one only for the entire low-dimensional search space $\mathbb{R}^d$, we show that it is sufficient to take specific compact sets. Our main outcome is to explicitly describe these minimal sets for searching a solution under the random embedding paradigm.

By pointing out to the difficulties that originate from the convex projection, we propose to alleviate these drawback by amending this component. In particular, we show that using an alternative embedding procedure yields a more convenient minimal set to work with, that is, relying on a back-projection from the orthogonal projection of the high-dimensional search space. We further show on examples that, in this case, the gain in robustness of discarding the risk of missing the optimum on the embedded low-dimensional space outweighs the increase in size of the search space.

The benefits could be even greater when extending the random embedding technique to constrained or multi-objective optimization, as tested e.g., in [1]. Indeed, the impact of restricting the search space too much could be even more important. Concerning Bayesian optimization, in addition to consider batch-sequential and other acquisition functions, perspectives include investigating non-stationary models to further improve the GP modeling aspect, based on the various properties uncovered. Finally, a promising approach would be to hybridize REMBO with methods that learn the low-dimensional structure as in [18].

Acknowledgments

We thank the anonymous reviewers for helpful comments on the earlier version of the paper. Parts of this work have been conducted within the frame of the ReDice Consortium, gathering industrial (CEA, EDF, IFPEN, IRSN, Renault) and academic (Ecole des Mines de Saint-Etienne, INRIA, and the University of Bern) partners around advanced methods for Computer Experiments. M. B. also acknowledges partial support from National Science Foundation grant DMS-1521702.

References

[1] Binois, M.: Uncertainty quantification on Pareto fronts and high-dimensional strategies in Bayesian optimization, with applications in multi-objective automotive design. Ph.D. thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne (2015)

[2] Binois, M., Ginsbourger, D., Roustant, O.: A warped kernel improving robustness in Bayesian optimization via random embeddings. In: C. Dhaenens, L. Jourdan, M.E. Marmion (eds.) Learning and Intelligent Optimization, *Lecture Notes in Com-*

puter Science, vol. 8994, pp. 281–286. Springer International Publishing (2015). DOI 10.1007/978-3-319-19084-6_28

[3] Carpentier, A., Munos, R.: Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In: International conference on Artificial Intelligence and Statistics (2012)

[4] Černỳ, M.: Goffin's algorithm for zonotopes. Kybernetika **48**(5), 890–906 (2012)

[5] Chen, B., Castro, R., Krause, A.: Joint optimization and variable selection of high-dimensional Gaussian processes. In: Proc. International Conference on Machine Learning (ICML) (2012)

[6] Chen, Y., Hoffman, M.W., Colmenarejo, S.G., Denil, M., Lillicrap, T.P., de Freitas, N.: Learning to learn for global optimization of black box functions. arXiv preprint arXiv:1611.03824 (2016)

[7] Constantine, P.G., Dow, E., Wang, Q.: Active subspace methods in theory and practice: applications to kriging surfaces. SIAM Journal on Scientific Computing **36**(4), A1500–A1524 (2014)

[8] Courrier, N., Boucard, P.A., Soulier, B.: Variable-fidelity modeling of structural analysis of assemblies. Journal of Global Optimization **64**(3), 577–613 (2016)

[9] Dixon, L., Szegö, G.: The global optimization problem: an introduction. Towards global optimization **2**, 1–15 (1978)

[10] Djolonga, J., Krause, A., Cevher, V.: High-dimensional Gaussian process bandits. In: Advances in Neural Information Processing Systems, pp. 1025–1033 (2013)

[11] Donoho, D.L.: High-dimensional data analysis: The curses and blessings of dimensionality. AMS Math Challenges Lecture pp. 1–32 (2000)

[12] Durrande, N.: Étude de classes de noyaux adaptées à la simplification et à linterprétation des modèles dapproximation. une approche fonctionnelle et probabiliste. Ph.D. thesis, Saint-Etienne, EMSE (2011)

[13] Durrande, N., Ginsbourger, D., Roustant, O.: Additive kernels for Gaussian process modeling. Annales de la Facultée de Sciences de Toulouse **21**(3), 481–499 (2012)

[14] Duvenaud, D.K.: Automatic model construction with Gaussian processes. Ph.D. thesis, University of Cambridge (2014)

[15] Feliot, P., Bect, J., Vazquez, E.: A Bayesian approach to constrained single-and multi-objective optimization. Journal of Global Optimization pp. 1–37 (2015)

[16] Filliman, P.: Extremum problems for zonotopes. Geometriae Dedicata **27**(3), 251–262 (1988)

[17] Franey, M., Ranjan, P., Chipman, H.: Branch and bound algorithms for maximizing expected improvement functions. Journal of Statistical Planning and Inference **141**(1), 42–55 (2011)

[18] Garnett, R., Osborne, M., Hennig, P.: Active learning of linear embeddings for Gaussian processes. In: 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014), pp. 230–239. AUAI Press (2014)

[19] Gutmann, H.M.: A radial basis function method for global optimization. Journal of Global Optimization **19**(3), 201–227 (2001)

[20] Hastie, T., Tibshirani, R., Friedman, J., Franklin, J.: The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer **27**(2), 83–85 (2005)

[21] Hennig, P., Schuler, C.J.: Entropy search for information-efficient global optimization. The Journal of Machine Learning Research **98888**, 1809–1837 (2012)

[22] Huang, D., Allen, T.T., Notz, W.I., Zeng, N.: Global optimization of stochastic black-box systems via sequential kriging meta-models. Journal of Global optimization **34**(3), 441–466 (2006)

[23] Hutter, F., Hoos, H.H., Leyton-Brown, K.: Sequential model-based optimization for general algorithm configuration. In: International Conference on Learning and Intelligent Optimization, pp. 507–523. Springer (2011)

[24] Iooss, B., Lemaître, P.: A review on global sensitivity analysis methods. In: C. Meloni, G. Dellino (eds.) Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications. Springer (2015)

[25] Ivanov, M., Kuhnt, S.: A parallel optimization algorithm based on FANOVA decomposition. Quality and Reliability Engineering International **30**(7), 961–974 (2014)

[26] Jones, D., Schonlau, M., Welch, W.: Efficient global optimization of expensive black-box functions. Journal of Global Optimization **13**(4), 455–492 (1998)

[27] Kandasamy, K., Schneider, J., Póczos, B.: High dimensional Bayesian optimisation and bandits via additive models. In: Proceedings of The 32nd International Conference on Machine Learning, pp. 295–304 (2015)

[28] Krein, M., Milman, D.: On extreme points of regular convex sets. Studia Mathematica **9**(1), 133–138 (1940)

[29] Krityakierne, T., Akhtar, T., Shoemaker, C.A.: SOP: parallel surrogate global optimization with Pareto center selection for computationally expensive single objective problems. Journal of Global Optimization pp. 1–21 (2016)

[30] Laguna, M., Martí, R.: Experimental testing of advanced scatter search designs for global optimization of multimodal functions. Journal of Global Optimization **33**(2), 235–255 (2005)

[31] Le, V.T.H., Stoica, C., Alamo, T., Camacho, E.F., Dumur, D.: Uncertainty representation based on set theory. Zonotopes pp. 1–26 (2013)

[32] Li, C.L., Kandasamy, K., Póczos, B., Schneider, J.: High dimensional bayesian optimization via restricted projection pursuit models. In: Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, pp. 884–892 (2016)

[33] Liu, B., Zhang, Q., Gielen, G.G.: A Gaussian process surrogate model assisted evolutionary algorithm for medium scale expensive optimization problems. Evolutionary Computation, IEEE Transactions on **18**(2), 180–192 (2014)

[34] Mardia, K.V., Kent, J.T., Bibby, J.M.: Multivariate Analysis (Probability and Mathematical Statistics). Academic Press (1980)

[35] Mathar, R., Zilinskas, A.: A class of test functions for global optimization. Journal of Global Optimization **5**(2), 195–199 (1994)

[36] McMullen, P.: On zonotopes. Transactions of the American Mathematical Society **159**, 91–109 (1971)

[37] Meyer, C.D.: Matrix analysis and applied linear algebra, vol. 2. Siam (2000)

[38] Mishra, S.: Global optimization by differential evolution and particle swarm methods: Evaluation on some benchmark functions. Tech. rep., University Library of Munich, Germany (2006)

[39] Mockus, J., Tiesis, V., Zilinskas, A.: The application of Bayesian methods for seeking the extremum. Towards Global Optimization **2**(117-129), 2 (1978)

[40] Morris, M.D., Mitchell, T.J., Ylvisaker, D.: Bayesian design and analysis of computer experiments: use of derivatives in surface prediction. Technometrics **35**(3), 243–255 (1993)

[41] Neal, R.M.: Bayesian learning for neural networks, *Lecture Notes in Statistics*, vol. 118. Springer (1996)

[42] Nguyen, H.H., Vu, V.: Random matrices: Law of the determinant. The Annals of Probability **42**(1), 146–167 (2014)

[43] Qian, H., Hu, Y.Q., Yu, Y.: Derivative-free optimization of high-dimensional non-convex functions by sequential random embeddings. In IJCAI 2016 (2016)

[44] Rasmussen, C.E., Williams, C.: Gaussian Processes for Machine Learning. MIT Press (2006)

[45] Rios, L.M., Sahinidis, N.V.: Derivative-free optimization: a review of algorithms and comparison of software implementations. Journal of Global Optimization **56**(3), 1247–1293 (2013)

[46] Roustant, O., Ginsbourger, D., Deville, Y.: DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. Journal of Statistical Software **51**(1), 1–55 (2012)

[47] Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., de Freitas, N.: Taking the human out of the loop: A review of Bayesian optimization. Proceedings of the IEEE **104**(1), 148–175 (2016)

[48] Song, W., Keane, A.J.: Surrogate-based aerodynamic shape optimization of a civil aircraft engine nacelle. AIAA journal **45**(10), 2565–2574 (2007)

[49] Turlach, B.A., Weingessel, A.: quadprog: Functions to solve Quadratic Programming Problems. (2013). URL https://CRAN.R-project.org/package=quadprog. R package version 1.5-5

[50] Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv:1011.3027 (2010)

[51] Villemonteix, J., Vazquez, E., Sidorkiewicz, M., Walter, E.: Global optimization of expensive-to-evaluate functions: an empirical comparison of two sampling criteria. Journal of Global Optimization **43**(2), 373–389 (2009)

[52] Viswanath, A., J. Forrester, A., Keane, A.: Dimension reduction for aerodynamic design optimization. AIAA journal **49**(6), 1256–1266 (2011)

[53] Wang, Z., Gehring, C., Kohli, P., Jegelka, S.: Batched large-scale bayesian optimization in high-dimensional spaces. In: International Conference on Artificial Intelligence and Statistics (2018)

[54] Wang, Z., Hutter, F., Zoghi, M., Matheson, D., de Feitas, N.: Bayesian optimization in a billion dimensions via random embeddings. Journal of Artificial Intelligence Research (JAIR) **55**, 361–387 (2016)

[55] Wang, Z., Zoghi, M., Hutter, F., Matheson, D., de Freitas, N.: Bayesian optimization in high dimensions via random embeddings. In IJCAI (2013)

[56] Ziegler, G.M.: Lectures on polytopes, vol. 152. Springer Science & Business Media (1995)

A Proofs

A.1 Properties of the convex projection

We begin with two elementary properties of the convex projection onto the hypercube $\mathcal{X} = [-1,1]^D$:

Property A.1 (Tensorization property)**.** $\forall \mathbf{x} \in \mathbb{R}^D$, $p_{\mathcal{X}} \begin{pmatrix} x_1 \\ \dots \\ x_D \end{pmatrix} = \begin{pmatrix} p_{[-1,1]}(x_1) \\ \dots \\ p_{[-1,1]}(x_D) \end{pmatrix}$.

Property A.2 (Commutativity with some isometries)**.** *Let q be an isometry represented by a diagonal matrix with terms $\varepsilon_i = \pm 1$, $1 \leq i \leq D$. Then, for all $\mathbf{x} \in \mathbb{R}^D$, $p_{\mathcal{X}}(q(\mathbf{x})) = \begin{pmatrix} \varepsilon_1 p_{[-1,1]}(x_1) \\ \dots \\ \varepsilon_D p_{[-1,1]}(x_D) \end{pmatrix} = q(p_{\mathcal{X}}(\mathbf{x}))$.*

A.2 Proof of Theorem 1

Proof. First, note that $\mathcal{U}$ is a closed set as a finite union of closed sets. Then, let us show that $p_{\mathcal{X}}(\mathbf{A}\mathcal{U}) = \mathcal{E}$. Consider $\mathbf{x} \in \mathcal{E}$, hence $|x_i| \leq 1$ and $\exists \mathbf{y} \in \mathbb{R}^d$ s.t. $\mathbf{x} = p_{\mathcal{X}}(\mathbf{A}\mathbf{y})$. Denote $\mathbf{b} = \mathbf{A}\mathbf{y}$. We distinguish two cases:

1. More than d components of $\mathbf{b}$ are in $[-1, 1]$. Then there exists a set $I \subset \{1, \dots, D\}$ of cardinality d such that $\mathbf{y} \in \bigcap_{i \in I} \mathcal{S}_i = \mathcal{P}_I \subseteq \mathcal{U}$, implying that $\mathbf{x} \in p_{\mathcal{X}}(\mathbf{A}\mathcal{U})$.

2. $0 \leq k < d$ components of $\mathbf{b}$ are in $[-1, 1]$. It is enough to consider that $\mathbf{b} \in [0, +\infty)^D$. Indeed, for any $\mathbf{x} \in \mathcal{E}$, any $\mathbf{A} \in \mathcal{A}$, let $\boldsymbol{\varepsilon}$ be the isometry given by the diagonal $D \times D$ matrix $\boldsymbol{\varepsilon}$ with elements ± 1 such that $\boldsymbol{\varepsilon}\mathbf{x} \in [0, +\infty)^D$. It follows that $\boldsymbol{\varepsilon}\mathbf{b}$ is in $[0, +\infty)^D$ too. Denote $\mathbf{x}' = \boldsymbol{\varepsilon}\mathbf{x}$, $\mathbf{b}' = \boldsymbol{\varepsilon}\mathbf{b}$ and $\mathbf{A}' = \boldsymbol{\varepsilon}\mathbf{A}$. Thus if $\exists \mathbf{u} \in \mathcal{U}$ such that $\mathbf{x}' = p_{\mathcal{X}}(\mathbf{b}') = p_{\mathcal{X}}(\mathbf{A}'\mathbf{u})$, by property A.2: $\boldsymbol{\varepsilon}\mathbf{x} = \boldsymbol{\varepsilon} p_{\mathcal{X}}(\mathbf{A}\mathbf{u})$ leading to $\mathbf{x} = p_{\mathcal{X}}(\mathbf{b}) = p_{\mathcal{X}}(\mathbf{A}\mathbf{u})$. From now on, we therefore assume that $b_i \geq 0$, $1 \leq i \leq D$. Furthermore, we can assume that $0 \leq b_1 \leq \cdots \leq b_D$, from a permutation of indices. Hence $b_i > 1$ if $i > k$ and $\mathbf{x} = (x_1 = b_1, \dots, x_k = b_k, 1, \dots, 1)^T$.

   Let $\mathbf{y}' \in \mathbb{R}^d$ be the solution of $\mathbf{A}_{1,\dots,d}\mathbf{y}' = (b_1, \dots b_k, 1, \dots, 1)^T$ (vector of size d). Such a solution exists since $\mathbf{A}_{1,\dots,d}$ is invertible by hypothesis. Then define $\mathbf{b}' = \mathbf{A}\mathbf{y}'$, $\mathbf{b}' = (b_1, \dots, b_k, 1, \dots, 1, b'_{d+1}, \dots, b'_D)^T$. We have $\mathbf{b}' \in \mathrm{Ran}(\mathbf{A})$ and $\mathbf{y}' \in \mathcal{P}_{1,\dots,d} \subseteq \mathcal{U}$.

   - If $\min_{i \in \{d+1,\dots,D\}}(b'_i) \geq 1$, then $p_{\mathcal{X}}(\mathbf{b}') = p_{\mathcal{X}}(\mathbf{b}) = \mathbf{x}$, and thus $\mathbf{x} = p_{\mathcal{X}}(\mathbf{A}\mathbf{y}') \in p_{\mathcal{X}}(\mathbf{A}\mathcal{U})$.
   - Else, the set $L = \{i \in \mathbb{N} : d + 1 \leq i \leq D\,,\, b'_i < 1\}$ is not empty. Consider $\mathbf{c} = \lambda\mathbf{b}' + (1 - \lambda)\mathbf{b}$, $\lambda \in]0, 1[$. By linearity, since both $\mathbf{b}$ and $\mathbf{b}'$ belong to $\mathrm{Ran}(\mathbf{A})$, $\mathbf{c} \in \mathrm{Ran}(\mathbf{A})$.
     - For $1 \leq i \leq k$, $c_i = x_i$.
     - For $k + 1 \leq i \leq d$, $c_i = \lambda + (1 - \lambda)b_i \geq 1$ since $b_i > 1$.
     - For $i \in \{d + 1, \dots, D\} \setminus L$, $b'_i \geq 1$ and $b_i > 1$ hence $c_i \geq 1$.
     - We now focus on the remaining components in L. For all $i \in L$, we solve $c_i = 1$, i.e., $\lambda b'_i + (1 - \lambda)b_i = \lambda(b'_i - b_i) + b_i = 1$. The solution is $\lambda_i = \frac{b_i - 1}{b_i - b'_i}$, with $b_i - b'_i > 0$ since $b'_i < 1$. Also $b_i - 1 > 0$ and $b_i - 1 < b_i - b'_i$ such that we have $\lambda_i \in]0, 1[$. Denote $\lambda^* = \min_{i \in L} \lambda_i$ and the corresponding index i^*. By construction, $c_{i^*} = 1$ and $\forall i \in L$, $c_i = \lambda^*(b'_i - b_i) + b_i \geq \lambda_i(b'_i - b_i) + b_i = 1$ since $\lambda_i \geq \lambda^*$ and $b'_i - b_i < 0$.

   To summarize, we can construct $\mathbf{c}^*$ with λ^* that has $k + 1$ components in $[-1, 1]$ (the first k and the i^{*th} ones), the others are greater or equal than 1. Moreover, $\mathbf{c}^* \in \mathrm{Ran}(\mathbf{A})$ and fulfills $p_{\mathcal{X}}(\mathbf{c}^*) = p_{\mathcal{X}}(\mathbf{b}) = \mathbf{x}$ by Property A.1. If $k + 1 = d$, this corresponds to case 1 above, otherwise, it is possible to reiterate by taking $\mathbf{b} = \mathbf{c}$. Hence we have a pre-image of $\mathbf{x}$ by ϕ in $\mathcal{U}$.

Thus the surjection property is shown. There remains to show that $\mathcal{U}$ is the smallest closed set achieving this, along with additional topological properties.

Let us show that any closed set $\mathcal{Y} \in \mathbb{R}^d$ such that $p_{\mathcal{X}}(\mathbf{A}\mathcal{Y}) = \mathcal{E}$ contains $\mathcal{U}$. To this end, we consider $\mathcal{U}^* = \bigcup_{I\subseteq\{1,\dots,D\},|I|=d} \mathring{\mathcal{P}}_I$ with $\mathring{\mathcal{P}}_I = \left\{\mathbf{y} \in \mathbb{R}^d\,, \forall i \in I, -1 < \mathbf{A}_i\mathbf{y} < 1\right\}$, the interior of the parallelotopes. We have $\phi|_{\mathring{U}}$ bijective. Indeed, all $\mathbf{x} \in p_{\mathcal{X}}(\mathbf{A}\mathcal{U}^*)$ have (at least) d components whose absolute value is strictly lower that 1. Without loss of generality, we suppose that they are the d first ones, $I = \{1, \dots, d\}$. Then there exists a *unique* $\mathbf{y} \in \mathbb{R}^d$ s.t. $\mathbf{x} = p_{\mathcal{X}}(\mathbf{A}\mathbf{y})$ because $\mathbf{x}_I = (\mathbf{A}\mathbf{y})_I = \mathbf{A}_I\mathbf{y}$ has a unique solution with $\mathbf{A}_I$ is invertible. Since $\mathcal{Y}$ is in surjection with $\mathcal{E}$ for $\phi|_{\mathcal{Y}}$ and $\phi|_{\mathcal{U}^*}$ is bijective, $\mathcal{U}^* \subseteq \mathcal{Y}$. Additionally, $\mathcal{Y}$ is a closed set so it must contain the closure $\mathcal{U}$ of $\mathcal{U}^*$.

Finally let us prove the topological properties of $\mathcal{U}$. Recall that parallelotopes $\mathcal{P}_I$ ($I \subseteq \{1, \dots, D\}$) are compact convex sets as linear transformations of d-cubes.
Thus $\mathcal{I} = \bigcap_{I\subseteq\{1,\dots,D\},|I|=d} \mathcal{P}_I$ is a compact convex set as the intersection of compact convex sets, which is non empty ($O \in \mathcal{I}$). It follows that $\mathcal{U} = \bigcup_{I\subseteq\{1,\dots,D\},|I|=d} \mathcal{P}_I$ is compact and connected as a finite union of compact connected sets with a non-empty intersection, i.e., $\mathcal{I}$. Additionally $\mathcal{U}$ is star-shaped with respect to any point in $\mathcal{I}$ (since $\mathcal{I}$ belongs to all parallelotopes in $\mathcal{U}$). □

A.3 Proof of Proposition 2.1

Proof. It follows from Definition 2.2 that $p_{\mathbf{A}}(\mathcal{X})$ is a zonotope of center O, obtained from the orthogonal projection of the D-hypercube $\mathcal{X}$. As such, $p_{\mathbf{A}}(\mathcal{X})$ is a convex polytope.

Since $\mathcal{E} \subset \mathcal{X}$, it is direct that $p_{\mathbf{A}}(\mathcal{E}) \subseteq p_{\mathbf{A}}(\mathcal{X})$.

To prove $p_{\mathbf{A}}(\mathcal{X}) \subseteq p_{\mathbf{A}}(\mathcal{E})$, let us start by vertices. Denote by $\mathbf{x} \in \mathbb{R}^D$ a vertex of $p_{\mathbf{A}}(\mathcal{X})$. If $\mathbf{x} \in \mathcal{X}$, then $p_{\mathbf{A}}(p_{\mathcal{X}}(\mathbf{x})) = p_{\mathbf{A}}(\mathbf{x}) = \mathbf{x}$, i.e., $\mathbf{x}$ has a pre-image in $\mathcal{E}$ by $p_{\mathbf{A}}$.
Else, if $\mathbf{x} \notin \mathcal{X}$, consider the vertex $\mathbf{v}$ of $\mathcal{X}$ such that $p_{\mathbf{A}}(\mathbf{v}) = \mathbf{x}$. Suppose that $\mathbf{v} \notin \mathcal{E}$. Let us remark that if $\mathbf{v}$ is a vertex of $\mathcal{X}$ such that $\mathbf{v} \notin \mathcal{E}$, then $\mathrm{Ran}(\mathbf{A}) \cap H_v = \emptyset$, where H_v is the open hyper-octant (with strict inequalities) that contains $\mathbf{v}$. Indeed, if $\mathbf{x} \in \mathrm{Ran}(\mathbf{A}) \cap H_v$, $\exists k \in \mathbb{R}^*$ such that $p_{\mathcal{X}}(k\mathbf{x}) = \mathbf{v}$, which contradicts $\mathbf{v} \notin \mathcal{E}$. Denote by $\mathbf{u}$ the intersection of the line $(O\mathbf{x})$ with $\mathcal{X}$, since $\mathbf{x} \notin H_v$, $\mathbf{u} \notin H_v$ either, hence $\widehat{\mathbf{xuv}} > \pi/2$. Then $\|\mathbf{u}-\mathbf{v}\| \leq \|\mathbf{x}-\mathbf{v}\|$, which contradicts $\mathbf{x} = p_{\mathbf{A}}(\mathbf{v})$. Hence $\mathbf{v} \in \mathcal{E}$ and $\mathbf{x}$ has a pre-image by $p_{\mathbf{A}}$ in $\mathcal{E}$.

Now, suppose $\exists \mathbf{x} \in p_{\mathbf{A}}(\mathcal{X})$ such that its pre-image(s) in $\mathcal{X}$ by $p_{\mathbf{A}}$ belong to $\mathcal{X} \setminus \mathcal{E}$. Denote $\mathbf{x}' \in p_{\mathbf{A}}(\mathcal{X})$ the closest vertex of $p_{\mathbf{A}}(\mathcal{X})$, which has a pre-image in $\mathcal{E}$ by $p_{\mathbf{A}}$. By continuity of $p_{\mathbf{A}}$, there exists $\mathbf{x}'' \in [\mathbf{x}, \mathbf{x}']$ with pre-image in $(\mathcal{X} \setminus \mathcal{E}) \cap \mathcal{E} = \emptyset$, hence there is a contradiction since $\mathbf{x}''$ has at least one pre-image. Consequently $\mathbf{x}$ has at least a pre-image in $\mathcal{E}$, and $p_{\mathbf{A}}(\mathcal{X}) \subseteq p_{\mathbf{A}}(\mathcal{E})$. □

A.4 Proof of Theorem 2

Proof. As a preliminary, let us show that $\forall \mathbf{y} \in \mathcal{Z}$, $\gamma(\mathbf{y}) \in \mathcal{E}$. Let $\mathbf{x}_1 \in \mathcal{X} \cap p_{\mathbf{A}}^{-1}(\mathbf{B}^\top\mathbf{y})(\neq \emptyset)$. From Proposition 2.1, $\mathbf{y}$ also have a pre-image $\mathbf{x}_2 \in \mathcal{E}$ by $p_{\mathbf{A}}$, and denote $\mathbf{u} \in \mathrm{Ran}(\mathbf{A})$ such that $p_{\mathcal{X}}(\mathbf{u}) = \mathbf{x}_2$, i.e., $\|\mathbf{x}_2-\mathbf{u}\| = \min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}-\mathbf{u}\|$. We have $\|\mathbf{x}_1-\mathbf{u}\|^2 = \|\mathbf{x}_1-\mathbf{B}^\top\mathbf{y}\|^2+\|\mathbf{B}^\top\mathbf{y}-\mathbf{u}\|^2$ and $\|\mathbf{x}_2-\mathbf{u}\|^2 = \|\mathbf{x}_2-\mathbf{B}^\top\mathbf{y}\|^2 + \|\mathbf{B}^\top\mathbf{y}-\mathbf{u}\|^2$ as $\mathbf{x}_1, \mathbf{x}_2 \in p_{\mathbf{A}}^{-1}(\mathbf{B}^\top\mathbf{y})$. Then, $\|\mathbf{x}_2-\mathbf{u}\| \leq \|\mathbf{x}_1-\mathbf{u}\| \Rightarrow \|\mathbf{x}_2-\mathbf{B}^\top\mathbf{y}\| \leq \|\mathbf{x}_1-\mathbf{B}^\top\mathbf{y}\|$ with equality if $\mathbf{x}_1 = \mathbf{x}_2$, so that $\gamma(\mathbf{B}^\top\mathbf{y}) \in \mathcal{E}$.

We now proceed by showing that γ defines a bijection from $\mathcal{Z}$ to $\mathcal{E}$, with $\gamma^{-1} = \mathbf{B}$. First, $\forall \mathbf{y} \in \mathcal{Z}$, $\mathbf{B}\gamma(\mathbf{y}) = \mathbf{y}$ by definition of γ. It remains to show that, $\forall \mathbf{x} \in \mathcal{E}$, $\gamma(\mathbf{B}\mathbf{x}) = \mathbf{x}$. Let

$\mathbf{x} \in \mathcal{E}$, $\mathbf{u} \in \mathrm{Ran}(\mathbf{A})$ such that $p_{\mathcal{X}}(\mathbf{u}) = \mathbf{x}$. Suppose $\gamma(\mathbf{Bx}) = \mathbf{x}' \in \mathcal{E}$, $\mathbf{x}' \neq \mathbf{x}$, in particular $\|\mathbf{x}' - \mathbf{B}^\top\mathbf{Bx}\| < \|\mathbf{x} - \mathbf{B}^\top\mathbf{Bx}\|$. Again, $\mathbf{x}, \mathbf{x}' \in p_{\mathbf{A}}^{-1}(\mathbf{B}^\top\mathbf{Bx})$, hence $\|\mathbf{x}' - \mathbf{B}^\top\mathbf{Bx}\|^2 + \|\mathbf{B}^\top\mathbf{Bx} - \mathbf{u}\|^2 = \|\mathbf{x}' - \mathbf{u}\|^2 < \|\mathbf{x} - \mathbf{B}^\top\mathbf{Bx}\|^2 + \|\mathbf{B}^\top\mathbf{Bx} - \mathbf{u}\|^2 = \|\mathbf{x} - \mathbf{u}\|^2$ which contradicts $\mathbf{x} = p_{\mathcal{X}}(\mathbf{u})$. Thus $\gamma(\mathbf{Bx}) = \mathbf{x}$.

$\mathcal{Z}$ is compact, convex and centrally symmetric from being a zonotope, see Definition 2.2. Finally, any smaller set than $\mathcal{Z}$ would not have an image through γ covering $\mathcal{E}$ entirely, which concludes the proof. □

A.5 Proof of Proposition 2.2

Proof. The first part directly follows from the properties of the convex and orthogonal projection. In detail: $\mathrm{Vol}_d(\mathbf{A}\mathcal{U}) \geq \mathrm{Vol}_d(p_{\mathcal{X}}(\mathbf{A}\mathcal{U})) = \mathrm{Vol}_d(\mathcal{E}) \geq \mathrm{Vol}_d(p_{\mathbf{A}}(\mathcal{E})) = \mathrm{Vol}_d(\mathbf{A}\mathcal{Z})$.

For the second part, we need the length of a strip $\mathcal{S}_i$, i.e., the inter hyperplane distance: $l_i = 2/\|\mathbf{A}_i\|$. Recall that $\mathbf{B} = \mathbf{A}^\top$, that rows of $\mathbf{A}$ have equal norm and $\mathbf{A}$ orthonormal. Then, following the proof of [16, Theorem 1], $\sum_{j=1}^{d} \|\mathbf{B}_j\|^2 = d$ (orthonormality) $= \sum_{i=1}^{D}\sum_{j=1}^{d} A_{i,j}^2 = \sum_{i=1}^{D} \|\mathbf{A}_i\|^2 = D\|\mathbf{A}_1\|^2$, hence $\|\mathbf{A}_1\| = \sqrt{d/D}$. As $\mathcal{Z}$ is enclosed in the d-sphere of radius $\sqrt{D}$ and the d-sphere of radius $\sqrt{D/d}$ is enclosed in $\mathcal{I}$, the result follows from the formula of the volume of a d-sphere of radius ρ: $\frac{\pi^{d/2}}{\Gamma(d/2+1)}\rho^d$. □

B Main notations

d	low embedding dimension
D	original dimension, $D \gg d$
$\mathbf{A}$	random embedding matrix of size $D \times d$
$\mathbf{B}$	transpose of $\mathbf{A}$ after orthonormalization
$\mathcal{X}$	search domain $[-1, 1]^D$
$\mathcal{Y}$	low dimensional optimization domain, in $\mathbb{R}^d$
$\mathcal{Z}$	zonotope $\mathbf{B}\mathcal{X}$
$p_{\mathcal{X}}$	convex projection onto $\mathcal{X}$
$p_{\mathbf{A}}$	orthogonal projection onto $\mathrm{Ran}(\mathbf{A})$
Ψ	warping function from $\mathbb{R}^d$ to $\mathrm{Ran}(\mathbf{A})$
ϕ	mapping from $\mathcal{Y} \subset \mathbb{R}^d$ to $\mathbb{R}^D$
γ	mapping from $\mathcal{Z} \subset \mathbb{R}^d$ to $\mathbb{R}^D$
$(\mathcal{R})$	optimization problem for REMBO with mapping ϕ
$(\mathcal{R}')$	optimization problem for REMBO with mapping γ
$(\mathcal{Q})$	minimal volume problem definition with mapping ϕ
$(\mathcal{Q}')$	minimal volume problem definition with mapping γ
$\mathcal{B}$	box enclosing $\mathcal{Z}$
$\mathcal{E}$	image of $\mathbb{R}^d$ by ϕ
$\mathcal{S}_i$	strip associated with the i^{th} row of $\mathbf{A}$
$\mathcal{I}$	intersection of all strips $\mathcal{S}_i$
$\mathcal{U}$	union of all intersection of d strips $\mathcal{S}_i$
$\mathcal{P}_I$	parallelotope associated with strips in the set I