Expected decrease for derivative-free algorithms using random subspaces

Warren Hare * Lindon Roberts † Clément W. Royer ‡

August 9, 2023

Abstract

Derivative-free algorithms seek the minimum of a given function based only on function values queried at appropriate points. Although these methods are widely used in practice, their performance is known to worsen as the problem dimension increases. Recent advances in developing randomized derivative-free techniques have tackled this issue by working in low-dimensional subspaces that are drawn at random in an iterative fashion. The connection between the dimension of these random subspaces and the algorithmic guarantees has yet to be fully understood.

In this paper, we develop an analysis for derivative-free algorithms (both direct-search and model-based approaches) employing random subspaces. Our results leverage linear local approximations of smooth functions to obtain understanding of the expected decrease achieved per function evaluation. Although the quantities of interest involve multidimensional integrals with no closed-form expression, a relative comparison for different subspace dimensions suggest that low dimension is preferable. Numerical computation of the quantities of interest confirm the benefit of operating in low-dimensional subspaces.

AMS Subject classification: 65K05, 90C56, 90C60.

1 Introduction

Derivative-free algorithms are designed to minimize a function using solely function value information. These methods are particularly valuable for optimizing

*Department of Mathematics, University of British Columbia, Kelowna, British Columbia, Canada. Hare's research is partially funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) grant RGPIN-2023-03555. ORCID 0000-0002-4240-3903 (warren.hare@ubc.ca).

†School of Mathematics and Statistics, University of Sydney, Camperdown NSW 2006, Australia. ORCID 0000-0001-6438-9703 (lindon.roberts@sydney.edu.au).

‡LAMSADE, CNRS, Université Paris Dauphine-PSL, Place du Maréchal de Lattre de Tassigny, 75016 Paris, France. Royer's research is partially funded by Agence Nationale de la Recherche through program ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). ORCID 0000-0003-2452-2172 (clement.royer@lamsade.dauphine.fr).

functions arising in complex engineering and learning models and, as such, have been applied in a diversity of fields [1, 6, 18]. However, classical derivative-free algorithms typically struggle to optimize functions with a large number of variables, as they must explore a large variable space without the guidance provided by derivatives. For these algorithms, the number of function evaluations that are used at each iteration can scale linearly with the problem dimension. As a result, the use of derivative-free algorithms has historically been restricted to problems having no more than a hundred variables.

To overcome this fundamental limitation, recent algorithmic proposals have relied on applying iterations in randomly chosen subspaces. For example, several derivative-free algorithms based on *direct-search methods* have been proposed that use opposite Gaussian directions (effectively a one-dimensional subspace) in various settings as a way to compute steps using no more than two function evaluations [2, 9, 21]. Another line of work considered directions uniformly distributed in the unit sphere [7, 12]. In that setting, it was shown that an almost-surely convergent algorithm could be designed by using only two function evaluations per iteration, with the best choice (both in terms of gradient approximation and practical performance) being to use opposite directions [12]. More recently, a generalized analysis showed that random subspaces of arbitrary dimension could be used to design globally convergent methods [22]. Similar ideas were proposed in the context of finite-difference estimates aiming at approximating directional derivatives [17, 16].

Model-based derivative-free algorithms, that operate by maintaining a model of the objective function, have also been revisited using random subspaces. A model-based trust-region algorithm was recently proposed in the context of nonlinear least squares [4], drawing on similar ideas for derivative-based algorithms [3, 23]. A randomized subspace trust-region method was subsequently developed for stochastic optimization [10]. We also note the use of sketching matrices within derivative-free trust-region methods as another setup in which random subspaces can be employed [19].

In the direct-search setting, empirical performance strongly suggested that using one-dimensional subspaces provided the best results [12, 22]. The conclusions were not as definitive in the model-based case, where quadratic models seemingly required sufficiently large subspaces to be built in [4] (but that implementation incorporated numerous extra heuristics), while model-based algorithms using linear interpolation proved efficient using very low dimensions [10]. Although convergence analysis often applies for subspaces of any sufficiently large—but still $O(1)$—dimension, it does not provide a clear understanding of the connection between subspace dimension and practical performance, nor why extremely low-dimensional spaces (e.g. 1 or 2) are good choices in practice.

In this paper, we examine expected decrease for derivative-free algorithms based on random subspaces. To our knowledge, our approach of quantifying the expected per-iteration and per-oracle-call objective decrease is a novel framework for studying the complexity of randomized methods for nonlinear optimization. Our approach allows us to provide information about *average-case* algorithm performance, instead of the more common worst-case performance

analysis typical in complexity analysis (e.g. [12, 4]).

By considering a general algorithmic framework, we are able to handle both direct-search and model-based strategies. By leveraging local linear approximations of the function to minimize, we express our problem in terms of linear functions, which facilitates the derivation of decrease guarantees in expectation. Our analysis shows that using low subspace dimension leads to the best possible objective decrease per function evaluation. Since evaluating the objective is often the computational bottleneck of derivative-free algorithms, such a result further motivates the use of randomized subspaces in these methods.

The remainder of this paper is structured as follows. The rest of this introductory section sets the notations and recalls some useful results about uniform distributions in subspaces. Section 2 provides a general algorithm template that covers both direct-search and model-based techniques. Section 3 is dedicated to analyzing direct-search methods based on random subspaces. The corresponding results for model-based methods are described in Section 4. Section 5 illustrates our theoretical findings with numerical experiments. Finally, we discuss extensions of our results in Section 6.

1.1 Notations and probability background

Throughout the paper, d and p will always denote integers greater than or equal to 1 with $p \leq d$. The Euclidean norm in $\mathbb{R}^d$ will be denoted by $\|\cdot\|$. The identity matrix in $\mathbb{R}^{d\times d}$ will be denoted by I_d. The unit sphere in $\mathbb{R}^d$ will be denoted by $\mathcal{S}^{d-1}$. The set of orthogonal $d \times d$ matrices will be denoted by $O(d)$. For $p \leq d$, the Stiefel manifold of $p \times d$ matrices with orthogonal columns in $\mathbb{R}^d$ will be denoted by $\mathcal{V}_{p,d} := \{X \in \mathbb{R}^{d\times p} : X^{\mathrm{T}}X = I_p\}$. Note that $\mathcal{V}_{1,d}$ corresponds to the unit sphere S^{d-1} while $\mathcal{V}_{d,d} = O(d)$.

Our main results will really heavily on uniform distributions within the Stiefel manifold. Key results about this distribution are gathered in the next lemma, and we omit the proofs as they can be found in reference textbooks on normed vector spaces [20, Section 1] and manifolds [5, Section 2.2].

Lemma 1.1. *For any integers $1 \leq p \leq d$, the following hold.*

(i) The uniform distribution on $V_{p,d}$ is uniquely defined.

(ii) If X follows a uniform distribution on $V_{p,d}$, then so does $Q_1 X Q_2^{\mathrm{T}}$ for any (possibly random) $Q_1 \in V_{d,d}$ and $Q_2 \in V_{p,p}$ independent of X.

(iii) If Q follows a uniform distribution on $V_{d,d}$, then so does Q^{T}.

(iv) We may construct $X \in V_{p,d}$ uniformly distributed by $X = Q_1 X_0 Q_2^{\mathrm{T}}$ for fixed $X_0 \in V_{p,d}$, and independent and uniformly drawn $Q_1 \in V_{d,d}$ and $Q_2 \in V_{p,p}$.

2 Framework for derivative-free algorithms in random subspaces

In this section, we present a general framework for a derivative-free algorithm that performs steps in randomly drawn subspaces. Section 2.1 discusses our main framework, while Section 2.2 gives two variations on the general method: one for direct-search methods and one for model-based methods. Section 2.3 then defines the quantities of interest for analyzing the algorithms.

2.1 General framework

Consider the minimization of a continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, where the derivative of f cannot be used for algorithmic purposes. A derivative-free algorithm is an iterative procedure that explores the variable space by querying f at finitely many points at every iteration in order to select the next iterate. In this paper, we are interested in derivative-free algorithms that produces such iterates by evaluating f in a subspace of dimension $p \leq d$ at every iteration, where this subspace is drawn randomly.

Algorithm 1 provides the general framework for our analysis. At each iteration, a random subspace is selected and one iteration of a given derivative-free method (`DFi`) is performed on that subspace. As our analysis focuses on expected decrease per iteration, we intentionally leave the stopping criterion and the step size update procedure undefined.

Algorithm 1 Derivative-free algorithm with random subspaces

procedure DFAwRS($f, x^0, \delta^0, p, \max_{fc}, \epsilon_{\texttt{stop}}$, `DFi`)
  % f: the objective function, $f : \mathbb{R}^d \mapsto \mathbb{R}$
  % x^0: the initial point, $x^0 \in \mathbb{R}^d$
  % δ^0: the initial step size parameter, $\delta^0 > 0$
  % p: the subspace dimension, $p \in \{1, 2, \ldots, n\}$
  % `DFi`: iteration of the chosen DFO algorithm, used on subspaces
  while stopping conditions not met **do**
    Randomly select a subspace of dimension p with orthonormal basis

$$B = \{b_1, b_2, ..., b_p\} \subseteq \mathbb{R}^d$$

    Define $f^k|_p : \mathbb{R}^p \to \mathbb{R}$ as $f^k|_p(z) = f(x^k + \sum_{i=1}^p z_i b_i)$
    Create z^* from the output of one iteration of `DFi` applied to $f^k|_p$ using initial point $z^0 = 0$ and step size δ^k
    Set $x^{k+1} = x^k + \sum_{i=1}^p z_i^* b_i$
    Select δ^{k+1} and increment $k \leftarrow k + 1$
  end while
end procedure

In order to draw a random subspace at every iteration, we randomly generate

a random orthonormal basis B for a p-dimensional subspace of $\mathbb{R}^d$. In the rest of the paper, we will assume that B is generated from the uniform distribution on the Stiefel manifold $V_{p,d}$, which amounts to taking the first p columns of a uniformly sampled matrix from $O(d)$ [11, Section 2]. More precisely, given $Q = [q_1 \ \cdots \ q_d] \in O(d)$ uniformly sampled from the Haar measure, we construct the basis B by letting $b_i = q_i$ for all $i = 1, \ldots, p$.

In the next section, we illustrate two variants on this method corresponding to the two main classes of derivative-free algorithms.

2.2 Direct-search and model-based variants

Our first instance of `DFi` corresponds to a (directional) direct-search iteration. In their basic form, direct-search schemes do not attempt to build an approximate gradient, but merely explore the space along suitably chosen directions. In a deterministic setting, these directions usually form a positive spanning set, so that one of them is close to the steepest descent direction [15]. Recent proposals in a probabilistic setting have replaced this requirement by random directions, with a particular interest for using directions belonging to a random subspace [12, 22]. We adopt a similar approach in Algorithm 2 that describes our direct-search iteration.

Algorithm 2 Direct-search iteration (`ds`)

1: **procedure** `ds`$(f|_p, z, \delta)$
2: % $f|_p$: the objective function, $f|_p : \mathbb{R}^p \mapsto \mathbb{R}$
3: % z: the incumbent solution, $z \in \mathbb{R}^p$
4: % δ: the step size parameter, $\delta > 0$
5: Consider the canonical basis $\{e_1, e_2, \ldots, e_p\}$ for $\mathbb{R}^p$
6: Return $z^* = \operatorname{argmin}\{f|_p(z + \delta u) : u \in \{\pm e_i\}_{i=1}^p \cup \{0\}\}$
7: **end procedure**

Note that we restrict ourselves to using coordinate directions in Algorithm 2. This is only to simplify presentation. Indeed, applying Lemma 1.1(ii), it is clear that using a random orthonormal basis of $\mathbb{R}^p$ will produce the same expected decrease. Note also that line 6 of Algorithm 2 states that complete polling is performed, that is we sample in all directions and return the best point that can be obtained. We will discuss how this algorithmic choice can be relaxed in Section 3.

Our second algorithmic variant corresponds to a model-based iteration, and consists in building a linear interpolation model of the function. To this end, we leverage the notion of a simplex gradient [1, Chapter 9], which we restate below in a format tailored to our setup.

Definition 2.1. *Consider the function $f|_p$ used in Algorithm 1. Let $V = \begin{bmatrix} v^1 & v^2 & \ldots & v^p \end{bmatrix}$ be an invertible matrix in $\mathbb{R}^{p \times p}$. For any $z \in \mathbb{R}^p$, the* simplex

gradient *of* $f|_p$ *at* z *based on* D *is defined by*

$$\nabla_S f|_p(z, V) = (V^{\mathrm{T}})^{-1} \begin{bmatrix} f|_p(z+v^1) - f|_p(z) \\ f|_p(z+v^2) - f|_p(z) \\ \vdots \\ f|_p(z+v^p) - f|_p(z) \end{bmatrix}. \tag{2.1}$$

The simplex gradient is used to construct a linear interpolation model of f, that can be used to produce a step from z [1, Chpt 9]. This observation is at the heart of model-based derivative-free algorithms, where other, more elaborate models can be employed. In Algorithm 3, we describe a trust-region model-based iteration based on a simplex gradient. This iteration computes a step that minimizes the model $u \mapsto \nabla_S f|_p(z, I_p)^{\mathrm{T}} u$ over a ball of radius δ centered at z, with I_p being the identity matrix in $\mathbb{R}^{p\times p}$. In that simple case, the minimizer can be found explicitly, yielding formula (2.2).

Algorithm 3 Model-based iteration (`mb`)

procedure `mb`$(f|_p, z^0, \delta^0)$
  % $f|_p$: the objective function, $f|_p : \mathbb{R}^p \mapsto \mathbb{R}$
  % z: the incumbent solution, $z \in \mathbb{R}^p$
  % δ: the trust region radius, $\delta > 0$
  Evaluate $f|_p(z)$ and $f|_p(z + \delta e_i)$ $(i = 1, 2, \ldots, p)$ to construct

$$u = -\frac{\nabla_S f|_p(z, \delta \mathtt{I_p})}{\|\nabla_S f|_p(z, \delta \mathtt{I_p})\|} \tag{2.2}$$

  Return $z^* = \operatorname{argmin}\{f|_p(z),\ f|_p(z+\delta u)\}$
end procedure

Similarly to Algorithm 2, Algorithm 3 employs the coordinate directions in order to simplify presentation.

2.3 Expected decrease guarantees

Derivative-free algorithms are commonly designed so as to drive the step size or trust-region parameter δ^k to zero as the algorithm unfolds. Consequently, providing guarantees associated to the linear *Taylor model* of the function around any given point leads to guarantees about decrease in function values. We present one such result in Proposition 2.2.

Proposition 2.2. *Suppose that* f *is continuously differentiable with* L*-Lipschitz continuous gradient. Consider the* k*th iteration of Algorithm 1, and suppose that we find a random unit direction* $u \in \mathbb{R}^d$ *such that*

$$\mathbb{E}\left[\nabla f(x^k)^{\mathrm{T}} u\right] \leq -\gamma < 0, \tag{2.3}$$

where the expectation is taken over the randomness in u. Then, for sufficiently small δ_k,

$$\mathbb{E}\left[f(x^k + \delta^k u) - f(x_k)\right] \leq -\frac{\gamma}{2}\delta^k, \tag{2.4}$$

where the expectation is taken over the randomness in u.

Proof. By Taylor expansion and Lipschitz continuity, one has

$$f(x^k+\delta^k u) \leq f(x_k)+\delta^k \nabla f(x^k)^{\mathrm{T}} u+\frac{L}{2}(\delta^k)^2\|u\|^2 = f(x^k)+\delta^k \nabla f(x^k)^{\mathrm{T}} u+\frac{L}{2}(\delta^k)^2.$$

Taking expectations with respect to the randomness in u leads to

$$\begin{aligned}\mathbb{E}\left[f(x^k + \delta^k u) - f(x_k)\right] &\leq \delta^k \,\mathbb{E}\left[\nabla f(x^k)^{\mathrm{T}} u\right] + \frac{L}{2}(\delta^k)^2 \\ &\leq -\gamma\delta^k + \frac{L}{2}(\delta^k)^2.\end{aligned}$$

As a result, (2.4) is satisfied as long as $\delta^k < \frac{1}{L}$. □

Considering the consequences of Proposition 2.2, in the rest of the paper, we focus on the function

$$f^{lin}(x) = g^{\mathrm{T}} x, \tag{2.5}$$

where $g \in \mathbb{R}^n$. Analyzing such functions is significantly easier than the general nonlinear case. In particular, note that $f^{lin}(x + \delta d) - f^{lin}(x) = \delta g^{\mathrm{T}} x$ for any pair of vectors. As a result, the function variation scales linearly with $\|g\|$ and δ. In addition, upon applying Algorithm 3, note that any simplex gradient (using a well-poised sample set) will always be equal to the actual gradient g, regardless of the value of δ [1, Exer 9.4]. Therefore, we also assume without loss of generality that $\delta = \|g\| = 1$.

We are interested in the expected decrease that one can achieve over one iteration of a derivative-free algorithm regardless of the value of g. This leads us to the following definition.

Definition 2.3. *Consider applying one iteration Algorithm 1 using either Algorithm 2 or Algorithm 3, denoted by* $\texttt{DFi} \in \{\texttt{ds}, \texttt{mb}\}$*, to a function $f^{lin}|_p$ obtained from f^{lin} defined in (2.5) with a vector g uniformly distributed on the unit sphere $\mathcal{S}^{d-1}$, using $\delta = 1$ and $p \leq n$. We define the* expected decrease $\mathbb{E}_{\texttt{DFi}}[p, d]$ *as*

$$\mathbb{E}_{\texttt{DFi}}[p, d] := \mathbb{E}\left[f^{lin}(x^k) - f^{lin}(x^{k+1})\right], \tag{2.6}$$

where the expected value is taken over g and B.

Our key results, presented in Sections 3 and 4, aim at providing formulae for the quantity (2.6). In both cases, we will see that B does not influence the value of the expected decrease.

3 Analysis in the direct-search setting

In this section, we examine the expected decrease for an iteration described by Algorithm 2. Our main result will be obtained in Section 3.1 using bounds on multidimensional integrals, and we will discuss consequences in Section 3.2 in terms of relative decrease per function evaluation.

3.1 Expected decrease formula

As a preliminary result, we show that the expected decrease produced by Algorithm 2 is independent of the random subspace basis B.

Proposition 3.1. *Consider the linear function f^{lin} with $g \sim \mathcal{S}^{d-1}$, and suppose that Algorithm 1 is applied using Algorithm 2 as* `DFi` *(which we denote by* `DFi`$=$`ds`*). Then, for any k, the expected decrease satisfies*

$$\mathbb{E}_{\texttt{ds}}[p,d] = \mathbb{E}_{\tilde{g}\sim\mathcal{S}^{d-1}}\left[\max_{i=1,\ldots,p}|\tilde{g}_i|\right]. \tag{3.1}$$

Proof. We first note that $x^k = x^{k+1}$ only when B is orthogonal to g, and this occurs with probability 0. Therefore, without loss of generality we assume that $x^k \neq x^{k+1}$ so that $f(x^k) - f(x^{k+1}) > 0$. In that case, letting $B = [b_1 \ \cdots \ b_p]$, we have

$$\begin{aligned} f(x^k) - f(x^{k+1}) &= \max_{i=1,\ldots,p} g^{\mathrm{T}}x^k - g^{\mathrm{T}}(x^k \pm b_i) \\ &= \max_{i=1,\ldots,p} |g^{\mathrm{T}}b_i| \\ &= \|B^{\mathrm{T}}g\|_\infty. \end{aligned}$$

As a result,

$$\mathbb{E}_{\texttt{ds}}[p,d] = \mathbb{E}_{\substack{B\sim\mathcal{V}_{p,d}\\ g\sim\mathcal{V}_{1,d}}}\left[\|B^{\mathrm{T}}g\|_\infty\right].$$

Let $I_{d,p} := [e_1,\ldots,e_p] \in \mathcal{V}_{p,d}$ be the matrix containing the first p coordinate directions in $\mathbb{R}^d$. By Lemma 1.1(iv), we have $B = QI_{d,p}$ for some $Q \sim \mathcal{V}_{d,d}$. Moreover, by Lemma 1.1(ii) and (iii), the random vector $Q^{\mathrm{T}}g$ follows the same distribution than g, i.e. uniform distribution in $\mathcal{V}_{1,d}$. Therefore, we obtain

$$\begin{aligned} \mathbb{E}_{\texttt{ds}}[p,d] &= \mathbb{E}_{\substack{B\sim\mathcal{V}_{p,d}\\ g\sim\mathcal{V}_{1,d}}}\left[\|B^{\mathrm{T}}g\|_\infty\right] \\ &= \mathbb{E}_{\substack{Q\sim\mathcal{V}_{d,d}\\ g\sim\mathcal{V}_{1,d}}}\left[\|I_{d,p}^{\mathrm{T}}Q^{\mathrm{T}}g\|_\infty\right] \\ &= \mathbb{E}_{\tilde{g}\sim\mathcal{V}_{1,d}}\left[\|I_{d,p}^{\mathrm{T}}\tilde{g}\|_\infty\right] \\ &= \mathbb{E}_{\tilde{g}\sim\mathcal{V}_{1,d}}\left[\max_{i=1,\ldots,p}|\tilde{g}_i|\right], \end{aligned}$$

proving (3.1). □

We will now obtain a mathematical expression for the expectation (3.1). When $d = 1$, we necessarily have $p = 1$ and

$$\mathbb{E}_{\mathtt{ds}}[1, 1] = 1.$$

Our main result will thus focus on the case $d > 1$. In general, the expected decrease formula is considerably more intricate, as it involves multiple Gamma functions as well as the solution to a complex trigonometry integral.

Theorem 3.2. *Under the assumptions of Proposition 3.1, suppose further that $d > 1$. Then, the expected decrease is given by*

$$\mathbb{E}_{\mathtt{ds}}[p, d] = \frac{p}{2} \frac{2^p}{(\sqrt{\pi})^p} \frac{\Gamma(d/2)\Gamma(p/2 + 1/2)}{\Gamma(d/2 + 1/2)} \mathcal{I}(p), \tag{3.2}$$

where $\mathcal{I}(p)$ is given by $\mathcal{I}(1) := 1$ and

$$\mathcal{I}(p) := \int_{R(p)} \left[\prod_{i=1}^{p-1} \sin^i(\varphi_i) \right] d\varphi_{p-1} \cdots d\varphi_1, \qquad \text{if } p > 1, \tag{3.3}$$

with the integration region $R(p)$ is $\{\varphi_1 \in [\pi/4, \pi/2]\}$ if $p = 1$ and

$$\left\{ (\varphi_1, \ldots, \varphi_{p-1}) \in [\pi/4, \pi/2] \times \prod_{i=2}^{p-1} \left[\arctan\left(\prod_{j=1}^{i-1} \csc \varphi_j \right), \frac{\pi}{2} \right] \right\} \tag{3.4}$$

otherwise.

Proof. By Proposition 3.1, we seek to evaluate the expectation (3.1), i.e.,

$$\mathbb{E}_{\mathtt{ds}}[p, d] = \mathbb{E}_{\tilde{g} \sim \mathcal{S}^{d-1}} \left[\max_{i=1,\ldots,p} |\tilde{g}_i| \right].$$

To this end, it suffices to evaluate the integral over the region

$$R(p, d) := \left\{ \tilde{g} \in \mathcal{S}^{d-1} \;\middle|\; \tilde{g}_1 \geq \tilde{g}_i \geq 0 \quad \forall i = 1, \ldots, p \right\},$$

i.e., vectors in the nonnegative orthant for which the first coordinate is the largest. By symmetry, one can construct $p2^d$ similar regions with the same integral value by selecting a maximal absolute value coordinate and an orthant. Thus, integrating over $R(p, d)$ gives $1/(p2^d)$ of the total integral. Moreover, for any $\tilde{g} \in R(p, d)$, we get the simplification

$$\max\{|\tilde{g}_1|, \ldots, |\tilde{g}_p|\} = \tilde{g}_1,$$

and therefore (3.1) can be rewritten as

$$\mathbb{E}_{\mathtt{ds}}[p, d] = \frac{p2^d}{|\mathcal{S}^{d-1}|} \int_{R(p,d)} \tilde{g}_1 \, dS(\tilde{g}), \tag{3.5}$$

where dS is the surface element for $\mathcal{S}^{d-1}$ and $|\mathcal{S}^{d-1}|$ is the volume of the unit sphere in $\mathbb{R}^d$.

To evaluate (3.5), we use hyperspherical coordinates $(\varphi_1, \ldots, \varphi_{d-1})$ for $\mathcal{S}^{d-1}$:

$$\begin{aligned} x_d &= \cos(\varphi_1), \\ x_{d-1} &= \sin(\varphi_1)\cos(\varphi_2), \\ &\vdots \\ x_2 &= \sin(\varphi_1)\cdots\sin(\varphi_{d-2})\cos(\varphi_{d-1}), \\ x_1 &= \sin(\varphi_1)\cdots\sin(\varphi_{d-2})\sin(\varphi_{d-1}), \end{aligned}$$

with surface element

$$dS = \sin^{d-2}(\varphi_1)\sin^{d-3}(\varphi_2)\cdots\sin(\varphi_{d-2})d\varphi_1 d\varphi_2\cdots d\varphi_{d-1}.$$

Note that this choice is a reverse of the traditional ordering of the axes, that will result in a simpler proof.

With these coordinates, the constraints defining the region $R(p,d) \subset \mathcal{S}^{d-1}$ translate into the following constraints on $(\varphi_1, \ldots, \varphi_{d-1})$:

- $\tilde{g}_i \geq 0$ yields $\varphi_i \in [0, \pi/2]$ for all $i = 1, \ldots, d-1$;
- $\tilde{g}_1 \geq \tilde{g}_2$ yields $\sin(\varphi_{d-1}) \geq \cos(\varphi_{d-1})$, which simplifies to $\varphi_{d-1} \geq \pi/4$;
- $\tilde{g}_1 \geq \tilde{g}_3$ yields $\sin(\varphi_{d-2})\sin(\varphi_{d-1}) \geq \cos(\varphi_{d-2})$, which simplifies to

$$\varphi_{d-2} \geq \arctan(\csc(\varphi_{d-1})).$$

By continuing the process, we obtain the following description of $R(p,d)$ when $p = 1$:

$$R(p,d) = \{\varphi_i \in [0, \pi/2] \qquad \forall i = 1, \ldots, d-1\}.$$

When $p \geq 2$, then $R(p,d)$ is defined via the constraints

$$\varphi_{d-1} \in [\pi/4, \pi/2], \tag{3.6a}$$

$$\varphi_{d-i} \in \left[\arctan\left(\prod_{j=1}^{i-1}\csc(\varphi_{d-j})\right), \frac{\pi}{2}\right], \qquad i = 2, \ldots, p-1, \tag{3.6b}$$

$$\varphi_i \in [0, \pi/2], \qquad i = 1, \ldots, d-p. \tag{3.6c}$$

Thus, returning to equation (3.5), we find that

$$\begin{aligned} \mathbb{E}_{\mathtt{ds}}[p,d] &= \frac{p2^d}{|\mathcal{S}^{d-1}|}\int_{R(p,d)} \tilde{g}_1\, dS(\tilde{g}), \\ &= \frac{p2^d}{|\mathcal{S}^{d-1}|}\int_{R(p,d)} \left(\prod_{i=1}^{d-1}\sin(\varphi_i)\right)\left(\prod_{i=1}^{d-2}\sin^{d-i-1}(\varphi_i)\right) d\varphi_1\cdots d\varphi_{n-1}, \\ &= \frac{p2^d}{|\mathcal{S}^{d-1}|}\int_{R(p,d)} \left(\prod_{i=1}^{d-1}\sin^{d-i}(\varphi_i)\right) d\varphi_1\cdots d\varphi_{d-1}. \end{aligned}$$

When $p = 1$, the integral is fully separable, and we obtain

$$\mathbb{E}_{\mathtt{ds}}[1, d] = \frac{2^d}{|\mathcal{S}^{d-1}|} \prod_{i=1}^{d-1} \left(\int_0^{\pi/2} \sin^{d-i}(\theta) d\theta \right). \tag{3.7}$$

When $p > 1$, we can factor out the integration with respect to $\varphi_1, \ldots, \varphi_{d-p} \in [0, \pi/2]$, yielding

$$\begin{aligned} \mathbb{E}_{\mathtt{ds}}[p, d] \quad &= \frac{p 2^d}{|\mathcal{S}^{d-1}|} \left[\prod_{i=1}^{d-p} \int_0^{\pi/2} \sin^{d-i}(\theta) d\theta \right] \cdot \\ &\int_{\hat{R}(p,d)} \left(\prod_{i=d-p+1}^{d-1} \sin^{d-i}(\varphi_i) \right) d\varphi_{d-p+1} \cdots d\varphi_{d-1}, \end{aligned} \tag{3.8}$$

where the reduced integration region $\hat{R}(p, d)$ is parameterized by inclusions (3.6a) and (3.6b) only. For simplicity, we now relabel the variables $\varphi_{d-i} \mapsto \varphi_i$ in the reduced integral over $\hat{R}(p, d)$ so as to get

$$\begin{aligned} &\int_{\hat{R}(p,d)} \left(\prod_{i=d-p+1}^{d-1} \sin^{d-i}(\varphi_i) \right) d\varphi_{d-p+1} \cdots d\varphi_{d-1} \\ &= \int_{R(p)} \left(\prod_{i=1}^{p-1} \sin^{i}(\varphi_i) \right) d\varphi_{p-1} \cdots d\varphi_1, \end{aligned} \tag{3.9}$$

where the integration region $R(p)$ is defined by $\varphi_1 \in [\pi/4, \pi/2]$ and (3.4). Combining (3.7) for $p = 1$ with (3.8) and (3.9) for $p > 1$, we obtain overall that

$$\mathbb{E}_{\mathtt{ds}}[p, d] = \frac{p 2^d}{|\mathcal{S}^{d-1}|} \left[\prod_{i=1}^{n-p} \int_0^{\pi/2} \sin^{d-i}(\theta) d\theta \right] \mathcal{I}(p), \tag{3.10}$$

where $\mathcal{I}(p)$ is defined in (3.3).

Finally, we can simplify (3.10) using the identity

$$\int_0^{\pi/2} \sin^{d-i} \theta d\theta = \frac{\sqrt{\pi}}{2} \frac{\Gamma((d-i)/2 + 1/2)}{\Gamma((d-i)/2 + 1)}, \tag{3.11}$$

for any $i = 1, \ldots, d - p$. We then obtain

$$\begin{aligned} \prod_{i=1}^{d-p} \int_0^{\pi/2} \sin^{d-i}(\theta) d\theta \quad &= \quad \frac{\pi^{(d-p)/2}}{2^{(d-p)}} \frac{\Gamma(d/2)}{\Gamma(d/2 + 1/2)} \frac{\Gamma(d/2 - 1/2)}{\Gamma(d/2)} \cdots \frac{\Gamma(p/2 + 1/2)}{\Gamma(p/2 + 1)}, \\ &= \quad \frac{\pi^{(d-p)/2}}{2^{(d-p)}} \frac{\Gamma(p/2 + 1/2)}{\Gamma(d/2 + 1/2)}. \end{aligned} \tag{3.12}$$

Finally, applying (3.12) and $|\mathcal{S}^{d-1}| = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ to (3.10), we arrive at

$$\mathbb{E}_{\mathtt{ds}}[p, d] = \frac{p}{2} \frac{2^p}{(\sqrt{\pi})^p} \frac{\Gamma(d/2)\Gamma(p/2 + 1/2)}{\Gamma(d/2 + 1/2)} \mathcal{I}(p),$$

which is the desired result. □

We observe that the expression (3.2) is separable in p and d. This property allows for simplified expressions for certain values of p, and also results in simplifications while comparing two pairs of values for (p, d) as only one of the two dimension varies. We summarize these observations in the corollary below.

Corollary 3.3. *Let d_1, d_2, p_1, p_2 be integers greater than or equal to 1 such that $\max\{p_1, p_2\} \leq \max\{d_1, d_2\}$. Then, the following properties hold:*

(i) $\mathbb{E}_{\mathtt{ds}}[1, d_1] = \frac{1}{\sqrt{\pi}} \frac{\Gamma(d_1/2)}{\Gamma(d_1/2+1/2)}$;

(ii) if $d_1 > 2$, *then* $\mathbb{E}_{\mathtt{ds}}[2, d_1] = \frac{\sqrt{2}}{\sqrt{\pi}} \frac{\Gamma(d_1/2)}{\Gamma(d_1/2+1/2)}$;

(iii) $\frac{\mathbb{E}_{\mathtt{ds}}[p_1,d_1]}{\mathbb{E}_{\mathtt{ds}}[p_2,d_1]} = \frac{\mathbb{E}_{\mathtt{ds}}[p_1,d_2]}{\mathbb{E}_{\mathtt{ds}}[p_2,d_2]}$;

(iv) $\frac{\mathbb{E}_{\mathtt{ds}}[p_1,d_1]}{\mathbb{E}_{\mathtt{ds}}[p_1,d_2]} = \frac{\mathbb{E}_{\mathtt{ds}}[p_2,d_1]}{\mathbb{E}_{\mathtt{ds}}[p_2,d_2]}$.

Proof. The proofs of (i) and (ii) follow directly from (3.2) by using $\mathcal{I}(1) = 1$, $\Gamma(1) = 1$, $\mathcal{I}(2) = \int_{\pi/4}^{\pi/2} \sin(\varphi_1) d\varphi_1 = 1/\sqrt{2}$, and $\Gamma(3/2) = \frac{\sqrt{\pi}}{2}$.

The proofs of (iii) and (iv) exploit the separability of the expression (3.2). For any pair (p, d) of integers greater than or equal to 1, define

$$E(p) = \frac{p}{2} \frac{2^p}{(\sqrt{\pi})^p} \Gamma(p/2 + 1/2) \mathcal{I}(p) \quad \text{and} \quad \hat{E}(d) = \frac{\Gamma(d/2)}{\Gamma(d/2 + 1/2)}$$

so that $\mathbb{E}_{\mathtt{ds}}[p, d] = E(p)\hat{E}(d)$. Then,

$$\frac{\mathbb{E}_{\mathtt{ds}}[p_1, d_1]}{\mathbb{E}_{\mathtt{ds}}[p_2, d_1]} = \frac{E(p_1)}{E(p_2)} = \frac{\mathbb{E}_{\mathtt{ds}}[p_1, d_2]}{\mathbb{E}_{\mathtt{ds}}[p_2, d_2]}, \tag{3.13}$$

proving (iii), and

$$\frac{\mathbb{E}_{\mathtt{ds}}[p_1, d_1]}{\mathbb{E}_{\mathtt{ds}}[p_1, d_2]} = \frac{\hat{E}(d_1)}{\hat{E}(d_2)} = \frac{\mathbb{E}_{\mathtt{ds}}[p_2, d_1]}{\mathbb{E}_{\mathtt{ds}}[p_2, d_2]},$$

proving (iv). □

Another consequence of the separable nature of the expression (3.2) is that the asymptotic behaviour of this quantity as $d \to \infty$ depends entirely on p. To establish this property, we rely on the following lemma.

Lemma 3.4. *Asymptotically,*

$$\frac{\Gamma(d/2)}{\Gamma(d/2 + 1/2)} \to \frac{\sqrt{2}}{\sqrt{d}} \text{ as } d \to \infty.$$

Proof. Gautschi's inequality [8, Eq. (5.6.4)] states

$$x^{1-s} < \frac{\Gamma(x+1)}{\Gamma(x+s)} < (x+1)^{1-s},$$

for all $x > 0$ and $s \in (0, 1)$. Setting $x = \frac{d}{2}$ and $s = \frac{1}{2}$ provides

$$\frac{\sqrt{d}}{\sqrt{2}} < \frac{\Gamma(d/2+1)}{\Gamma(d/2+1/2)} < \frac{\sqrt{d+2}}{\sqrt{2}}. \tag{3.14}$$

Applying $\Gamma(d/2+1) = \frac{d}{2}\Gamma(d/2)$ now shows

$$\frac{\sqrt{2}}{\sqrt{d}} < \frac{\Gamma(d/2)}{\Gamma(d/2+1/2)} < \frac{\sqrt{2}\sqrt{d+2}}{d}.$$

Passing to a limit provides the asymptotic. □

Combining the result of Lemma 3.4 with Corollary 3.3(i) and (ii) leads to the following asymptotics.

Corollary 3.5. *Under the same assumptions as Corollary 3.3, asymptotically*

$$\mathbb{E}_{\mathtt{ds}}[1, d_1] \to \frac{\sqrt{2}}{\sqrt{\pi}\sqrt{d_1}} \text{ as } d_1 \to \infty$$

and for $d_1 > 2$, asymptotically

$$\mathbb{E}_{\mathtt{ds}}[2, d_1] \to \frac{2}{\sqrt{\pi}\sqrt{d_1}} \text{ as } d_1 \to \infty.$$

3.2 Expected decrease per function evaluation

Derivation-free algorithms are typically used in situations where function evaluations are considered to be expensive calculations. As such, the effectiveness of a derivative-free algorithm is not gauged by expected decrease per iteration, but expected decrease per function evaluation. We thus wish to account for this cost in our formula for expected decrease.

Returning to Algorithm 2, we assume that the function value of the incumbent solution x^k is already known from the output of the previous iteration. As such, one iteration of Algorithm 2 will evaluate the function at $2p$ new points, where p is the subspace dimension. In this section, we are thus interested in the quantity

$$\mathbb{E}^F_{\mathtt{ds}}[p, d] := \frac{\mathbb{E}_{\mathtt{ds}}[p, d]}{2p}. \tag{3.15}$$

Our goal is then to study the variation of the quantity $\mathbb{E}^F_{\mathtt{ds}}[p, d]$ as a function of p. In order to derive such a result, we require the following lemma.

Lemma 3.6. *Let the assumptions of Theorem 3.2 hold, and $\mathcal{I}(p)$ be defined as in this theorem. Then, for any $p \leq d-1$,*

$$\frac{2}{\sqrt{\pi}} \frac{\Gamma(p/2+1)}{\Gamma(p/2+1/2)} < \frac{\mathcal{I}(p)}{\mathcal{I}(p+1)}.$$

Proof. If $p = 1$, using the values of $\mathcal{I}(1)$ and $\mathcal{I}(2)$ from the proof of Corollary 3.3 gives

$$\mathcal{I}(p+1) = \mathcal{I}(2) = \frac{1}{\sqrt{2}} < 1 = \frac{\sqrt{\pi}}{2}\frac{\Gamma(1)}{\Gamma(3/2)}\mathcal{I}(1) = \frac{\sqrt{\pi}}{2}\frac{\Gamma(p/2+1/2)}{\Gamma(p/2+1)}\mathcal{I}(p).$$

Suppose now that $p > 1$. Using the definition of $\mathcal{I}(p)$ (3.3), we have

$$\mathcal{I}(p+1) = \int_{R(p)} \int_{\arctan(\csc(\varphi_1)\cdots\csc(\varphi_p))}^{\pi/2} \left[\prod_{i=1}^{p} \sin^i(\varphi_i)\right] d\varphi_p d\varphi_{p-1}\cdots d\varphi_1, \quad (3.16)$$

showing that $\mathcal{I}(p+1)$ is formed by including an extra inner integral inside the expression for $\mathcal{I}(p)$. We now bound the lower limit of region of integration for φ_p through induction. From the definition of $R(p)$, we have $\varphi_1 \in [\pi/4, \pi/2]$. Inductively, suppose that the region of integration of φ_i is a subset of $[\pi/4, \pi/2]$ for $i = 1, \ldots, k-1$. In that case, we have $\csc(\varphi_i) \in [1, \sqrt{2}]$ for all $i = 1, \ldots, k-1$ and so

$$\varphi_k \geq \arctan(\csc(\varphi_1)\cdots\csc(\varphi_{k-1})) \geq \arctan(1) = \pi/4,$$

which implies that the region of integration for φ_k is a subset of $[\pi/4, \pi/2]$.

By the principles of mathematical induction, we have thus established that the region of integration of φ_p is a subset of $[\pi/4, \pi/2]$. Applying this to the region of integration approximation to (3.16), we find

$$\begin{aligned}
\mathcal{I}(p+1) &\leq \int_{R(p)} \int_{\pi/4}^{\pi/2} \left[\prod_{i=1}^{p} \sin^i(\varphi_i)\right] d\varphi_p d\varphi_{p-1}\cdots d\varphi_1, \\
&= \left(\int_{R(p)} \left[\prod_{i=1}^{p-1} \sin^i(\varphi_i)\right] d\varphi_{p-1}\cdots d\varphi_1\right)\left(\int_{\pi/4}^{\pi/2} \sin^p(\varphi_p) d\varphi_p\right), \\
&= \mathcal{I}(p)\int_{\pi/4}^{\pi/2} \sin^p(\theta) d\theta.
\end{aligned}$$

The result now follows from

$$\int_{\pi/4}^{\pi/2} \sin^p(\theta) d\theta < \int_{0}^{\pi/2} \sin^p(\theta) d\theta = \frac{\sqrt{\pi}}{2}\frac{\Gamma(p/2+1/2)}{\Gamma(p/2+1)},$$

where the last equality uses the identity (3.16). □

Using the previous result, we can approximate the rate at which $\mathcal{I}$ decreases as a function of p.

Proposition 3.7. *Let the assumptions of Theorem 3.2 hold, and $\mathcal{I}(p)$ be defined as in Theorem 3.2. Then, for any $p \leq d-1$,*

$$\mathcal{I}(p+1) < \frac{\sqrt{\pi}}{\sqrt{2}\sqrt{p}}\mathcal{I}(p).$$

Proof. By Gautschi's inequality (see equation (3.14)), we have that

$$\frac{\Gamma(p/2+1/2)}{\Gamma(p/2+1)} < \frac{\sqrt{2}}{\sqrt{p}}.$$

Combining this with Lemma 3.6 completes the proof. □

We can now prove that the expected decrease per function evaluation is a strictly decreasing function of p.

Theorem 3.8. *Let the assumptions of Theorem 3.2 hold, and $\mathcal{I}(p)$ be defined as Theorem 3.2. Then, for any $p \leq d-1$,*

$$\frac{\mathbb{E}_{\mathtt{ds}}[p,d]}{2p} > \frac{\mathbb{E}_{\mathtt{ds}}[p+1,d]}{2(p+1)}$$

Proof. It suffices to show that

$$\frac{\mathbb{E}_{\mathtt{ds}}[p,d]}{\mathbb{E}_{\mathtt{ds}}[p+1,d]} > \frac{p}{p+1}.$$

As in the proof of Corollary 3.3, we define

$$E_1(p) = \frac{p}{2}\frac{2^p}{(\sqrt{\pi})^p}\Gamma(p/2+1/2)\mathcal{I}(p).$$

From equation (3.13), we have

$$\begin{aligned}
\frac{\mathbb{E}_{\mathtt{ds}}[p,d]}{\mathbb{E}_{\mathtt{ds}}[p+1,d]} &= \frac{E_1(p)}{E_1(p+1)} \\
&= \frac{(p/2)(2^p/\sqrt{\pi}^p)\Gamma(p/2+1/2)\mathcal{I}(p)}{(p/2+1/2)(2^{p+1}/\sqrt{\pi}^{p+1})\Gamma(p/2+1)\mathcal{I}(p+1)} \\
&= \frac{p}{p+1}\frac{\sqrt{\pi}}{2}\frac{\Gamma(p/2+1/2)}{\Gamma(p/2+1)}\frac{\mathcal{I}(p)}{\mathcal{I}(p+1)} \\
&> \frac{p}{p+1}\frac{\sqrt{\pi}}{2}\frac{\Gamma(p/2+1/2)}{\Gamma(p/2+1)}\frac{2}{\sqrt{\pi}}\frac{\Gamma(p/2+1)}{\Gamma(p/2+1/2)} \\
&= \frac{p}{p+1},
\end{aligned}$$

where the strict inequality arises from applying Lemma 3.6. □

The result of Theorem 3.8 suggests that performing direct-search iterations is more beneficial with low-dimensional subspaces, and that $p = 1$ provides the best return on investment. Although our result applies to a linear function, we emphasize again that it can be connected to general smooth functions through arguments such as that of Proposition 2.2.

To end this section, we discuss how our analysis can be adapted to classical considerations for direct-search methods in practice.

Opportunistic polling: In Algorithm 2, all candidate points are sampled in order to select the best one, i.e., complete polling is performed. In a serial environment, a cheaper practice called opportunistic polling consists in accepting the first point that yields decrease. Remarkably, this strategy does not jeopardize convergence and can bring significant savings in practice [15][22].

Our analysis can be adapted to account for opportunistic polling under the assumption that Algorithm 2 evaluates the directions in the order $\{e_1, -e_1, \ldots\}$ (or more generally by evaluating pairs of opposite directions consecutively). In that case, with probability 1 either e_1 or $-e_1$ will lead to a decrease in f^{lin}, and thus the step will be accepted. The expected decrease guarantees are therefore equivalent to those in the case $p = 1$. In fact, one can go one step further by considering that on average, one performs $3/2$ evaluations as e_1 has a 50% chance of being a direction of decrease. With that consideration, the expected decrease guarantee becomes

$$\frac{2}{3\sqrt{\pi}} \frac{\Gamma(d/2)}{\Gamma(d/2 + 1/2)},$$

which improves over the quantity (3.15) for $p = 1$. As this result even holds for $p = d$, this provides a novel explanation for the performance of direct-search approaches using opportunistic polling (with or without random subspaces).

Parallel processing: Using multiple cores to perform function evaluations in parallel is a common paradigm that affects the per-iteration workload. If c cores are dedicated to distinct function evaluations and complete polling is performed, then one can consider that Algorithm 2 has an evaluation cost of $\lceil 2p/c \rceil$, where $\lceil \cdot \rceil$ denotes the ceiling function. Since the expected decrease $\mathbb{E}_{\mathtt{ds}}[p, d]$ is a decreasing function of p (see, Corollary 3.3) and assuming $2p/c$ is an integer number, Theorem 3.8 implies that

$$\frac{\mathbb{E}_{\mathtt{ds}}[p, d]}{\lceil 2p/c \rceil} \geq c \frac{\mathbb{E}_{\mathtt{ds}}[p, d]}{2p} > c \frac{\mathbb{E}_{\mathtt{ds}}[p + c/2, d]}{2(p + c/2)} = \frac{\mathbb{E}_{\mathtt{ds}}[p + c/2, d]}{\lceil 2(p + c/2)/c \rceil}.$$

Consequently, in this parallel setting, the expected decrease per unit of work is maximized for $p = c/2$, i.e. the smallest subspace dimension that exploits all c cores. Such a result shows that our analysis can be adapted to the computational power available to perform function evaluations.

4 Analysis in the model-based setting

In this section, we examine expected decrease for Algorithm 3, i.e., when a model-based strategy is used to perform steps in the random subspace. The analysis is similar to that of Section 3 yet presents significant differences, as we will discuss below. Section 4.1 establishes the main expected decrease result, while Section 4.2 considers the results in light of per-iteration evaluation cost.

4.1 Expected decrease formula

We begin by deriving an expression for the expected decrease that does not depend on the selected basis for the random subspace.

Proposition 4.1. *Consider the linear function f^{lin} with $g \sim \mathcal{S}^{d-1}$, and suppose that Algorithm 1 is applied using Algorithm 3 as* `DFi` *(which we denote by* `DFi`$=$`mb`*) with $\delta^k = 1$. Then, for any k, the expected decrease guarantee satisfies*

$$\mathbb{E}_{\texttt{mb}}[p,d] = \mathbb{E}_{\tilde{g}\sim\mathcal{S}^{d-1}}\left[\sqrt{\sum_{i=1,\ldots,p}\tilde{g}_i^2}\right]. \tag{4.1}$$

Proof. As in the proof of Proposition 3.1, we assume without loss of generality that $x^{k+1} \neq x^k$. Let $B = [b_1 \cdots b_p]$ with $b_i \in \mathbb{R}^d$. Since $\delta^k = 1$, the simplex gradient calculated by Algorithm 3 is given by

$$\nabla_S f^{lin}|_p(x^k, \mathtt{I_p}) = I_{p\times p}\begin{bmatrix} f^{lin}(x^k+b_1) - f(x^k) \\ \vdots \\ f^{lin}(x^k+b_p) - f(x^k)\end{bmatrix} = B^{\mathrm{T}}g.$$

Therefore, the decrease obtained for $\delta^k = 1$ is

$$\begin{aligned} f(x^k) - f(x^{k+1}) &= f(x^k) - f\left(x^k - B\frac{B^{\mathrm{T}}g}{\|B^{\mathrm{T}}g\|}\right) \\ &= g^{\mathrm{T}}B\frac{B^{\mathrm{T}}g}{\|B^{\mathrm{T}}g} = \|B^{\mathrm{T}}g\|. \end{aligned}$$

In terms of expected decrease, we therefore obtain

$$\mathbb{E}_{\texttt{mb}}[p,d] = \mathbb{E}_{\substack{B\sim\mathcal{V}_{p,d}\\ g\sim\mathcal{V}_{1,d}}}\left[\|B^{\mathrm{T}}g\|\right].$$

By the same argument as in the proof of Proposition 4.1, we can write $B = QI_{d,p}$ with $Q \sim \mathcal{V}_{d,d}$ and $I_{d,p}$ containing the first p coordinate directions in $\mathbb{R}^d$, and $Q^{\mathrm{T}}g$ is uniformly distributed in $\mathcal{V}_{1,d}$. This leads to

$$\begin{aligned} \mathbb{E}_{\texttt{mb}}[p,d] &= \mathbb{E}_{\substack{B\sim\mathcal{V}_{p,d}\\ g\sim\mathcal{V}_{1,d}}}\left[\|B^{\mathrm{T}}g\|\right] \\ &= \mathbb{E}_{\tilde{g}\sim\mathcal{V}_{1,d}}\left[\|I_{d,p}^{\mathrm{T}}\tilde{g}\|\right] \\ &= \mathbb{E}_{\tilde{g}\sim\mathcal{V}_{1,d}}\left[\sqrt{\sum_{i=1,\ldots,p}\tilde{g}_i^2}\right], \end{aligned}$$

proving (4.1). □

We now derive an expression for (4.1). Similarly to the direct-search case, when $d = p = 1$, the expected decrease has a trivial expression

$$\mathbb{E}_{\texttt{mb}}[1,1] = 1.$$

We assume in the rest of this section that $d > 1$. In that case, the general form of the expected decrease is surprisingly elegant in that it does not include a trigonometric integral.

Theorem 4.2. *Under the assumptions of Proposition 4.1, suppose further that $d > 1$. Then, the expected decrease is given by*

$$\mathbb{E}_{\mathtt{mb}}[p, d] = \frac{\Gamma(d/2)\ \Gamma(p/2 + 1/2)}{\Gamma(d/2 + 1/2)\ \Gamma(p/2)}. \tag{4.2}$$

Proof. Our goal consists in evaluating the expression (4.1), i.e.

$$\mathbb{E}_{\mathtt{mb}}[p, d] = \mathbb{E}_{\tilde{g} \sim \mathcal{V}_{1,n}} \left[\sqrt{\sum_{i=1,\dots,p} \tilde{g}_i^2} \right].$$

Consider first the case $p = d$. Since $\tilde{g} \in \mathcal{S}^{d-1}$, we have $\sqrt{\sum_{i=1}^{p} \tilde{g}_i^2} = \|\tilde{g}\| = 1$, and thus

$$\mathbb{E}_{\mathtt{mb}}[d, d] = \mathbb{E}_{\tilde{g} \sim \mathcal{V}_{1,d}} [1] = 1.$$

Noting that formula (4.2) also returns 1 when $p = d$ shows that it is valid in that case. Thus, in the rest of the proof, we suppose that $p < d$.

In order to compute the expectation, we restrict ourselves to vectors in the nonnegative orthant, i.e. we consider $R := \{\tilde{g} \in \mathcal{S}^{d-1} \mid \tilde{g}_i \geq 0\ \forall i = 1, \dots, d\}$. As in the proof of Theorem 3.2, we introduce hyperspherical coordinates

$$\begin{aligned}
x_d &= \cos(\varphi_1), \\
x_{d-1} &= \sin(\varphi_1)\cos(\varphi_2), \\
&\vdots \\
x_2 &= \sin(\varphi_1)\cdots\sin(\varphi_{d-2})\cos(\varphi_{d-1}), \\
x_1 &= \sin(\varphi_1)\cdots\sin(\varphi_{d-2})\sin(\varphi_{d-1}),
\end{aligned}$$

with surface element

$$dS = \sin^{d-2}(\varphi_1)\sin^{d-3}(\varphi_2)\cdots\sin(\varphi_{d-2}) d\varphi_1 d\varphi_2 \cdots d\varphi_{d-1}.$$

(As before, we use the reverse of the traditional ordering in order to create a simpler proof.) Then, for any $\tilde{g}$ in the nonnegative orthant, we have

$$\sqrt{\sum_{i=1}^{p} \tilde{g}_i^2} = \prod_{k=1}^{d-p} \sin(\varphi_k).$$

Given that there are 2^d orthants in R^d, we obtain by symmetry that

$$\mathbb{E}_{\mathtt{mb}}[p, d] = \frac{2^d}{|\mathcal{S}^{d-1}|} \int_R \left(\prod_{k=1}^{d-p} \sin(\varphi_k) \right) \left(\prod_{k=1}^{d-2} \sin^{d-k-1}(\varphi_k) \right) d\varphi_1 \cdots d\varphi_{d-1},$$

where $|\mathcal{S}^{d-1}|$ denotes the volume of the unit sphere in $\mathbb{R}^d$. By exploiting partial separability of this integral, we obtain

$$\begin{aligned}
\mathbb{E}_{\mathtt{mb}}[p,d] &= \frac{2^d}{|\mathcal{S}^{d-1}|}\int_R \left(\prod_{k=1}^{d-p}\sin(\varphi_k)\right)\left(\prod_{k=1}^{d-2}\sin^{d-k-1}(\varphi_k)\right)d\varphi_1\cdots d\varphi_{d-1},\\
&= \frac{2^d\ \pi}{2|\mathcal{S}^{d-1}|}\int_R \left(\prod_{k=1}^{d-p}\sin(\varphi_k)\right)\left(\prod_{k=1}^{d-2}\sin^{d-k-1}(\varphi_k)\right)d\varphi_1\cdots d\varphi_{d-2},\\
&= \frac{2^d\ \pi}{2|\mathcal{S}^{d-1}|}\left(\prod_{k=1}^{d-p}\int_0^{\pi/2}\sin^{d-k}(\theta)d\theta\right)\left(\prod_{k=d-p+1}^{d-2}\int_0^{\pi/2}\sin^{d-k-1}(\theta)d\theta\right).
\end{aligned} \tag{4.3}$$

Recalling identity (3.11), we compute

$$\begin{aligned}
\prod_{k=d-p+1}^{d-2}\int_0^{\pi/2}\sin^{d-k-1}(\theta)d\theta &= \prod_{k=1}^{p-2}\int_0^{\pi/2}\sin^{p-k-1}(\theta)d\theta\\
&= \prod_{k=1}^{p-2}\int_0^{\pi/2}\sin^{k}(\theta)d\theta,\\
&= \prod_{k=1}^{p-2}\frac{\sqrt{\pi}\ \Gamma(k/2+1/2)}{2\ \Gamma(k/2+1)},\\
&= \left(\frac{\sqrt{\pi}}{2}\right)^{p-2}\frac{1}{\Gamma(p/2)}.
\end{aligned}$$

Also recalling (3.12) and substituting both into (4.3), we find that

$$\begin{aligned}
\mathbb{E}_{\mathtt{mb}}[p,d] &= \frac{2^d\ \pi}{2|\mathcal{S}^{d-1}|}\left(\frac{\sqrt{\pi}}{2}\right)^{d-p}\frac{\Gamma(p/2+1/2)}{\Gamma(d/2+1/2)}\left(\frac{\sqrt{\pi}}{2}\right)^{p-2}\frac{1}{\Gamma(p/2)}\\
&= \frac{2\ \pi^{d/2}}{|\mathcal{S}^{d-1}|}\frac{\Gamma(p/2+1/2)}{\Gamma(d/2+1/2)}\frac{1}{\Gamma(p/2)}\\
&= \frac{\Gamma(p/2+1/2)}{\Gamma(d/2+1/2)}\frac{\Gamma(d/2)}{\Gamma(p/2)},
\end{aligned}$$

where the final line comes from the substitution $|\mathcal{S}^{d-1}| = 2\pi^{d/2}/\Gamma(d/2)$. We have thus proved that (4.2) also holds in the case $p < d$, and the proof is complete. □

We examine several particular properties of the expression (4.2) in the next corollary. As in Section 3.1, we leverage the fact that the expression (4.2) has a separable structure.

Corollary 4.3. *Let d_1, d_2, p_1, p_2 be integers greater than or equal to 1 such that $\max\{p_1,p_2\} \le \max\{d_1,d_2\}$. Then, the following properties hold:*

(i) $\mathbb{E}_{\mathtt{mb}}[1, d_1] = \frac{1}{\sqrt{\pi}} \frac{\Gamma(d_1/2)}{\Gamma(d_1/2+1/2)}$;

(ii) *if* $d_1 > 2$, *then* $\mathbb{E}_{\mathtt{mb}}[2, d_1] = \frac{\sqrt{\pi}}{2} \frac{\Gamma(d_1/2)}{\Gamma(d_1/2+1/2)}$;

(iii) $\frac{\mathbb{E}_{\mathtt{mb}}[p_1,d_1]}{\mathbb{E}_{\mathtt{mb}}[p_2,d_1]} = \frac{\mathbb{E}_{\mathtt{mb}}[p_1,d_2]}{\mathbb{E}_{\mathtt{mb}}[p_2,d_2]}$;

(iv) $\frac{\mathbb{E}_{\mathtt{mb}}[p_1,d_1]}{\mathbb{E}_{\mathtt{mb}}[p_1,d_2]} = \frac{\mathbb{E}_{\mathtt{mb}}[p_2,d_1]}{\mathbb{E}_{\mathtt{mb}}[p_2,d_2]}$.

Notice that $\mathbb{E}_{\mathtt{mb}}[1, d] = \mathbb{E}_{\mathtt{ds}}[1, d]$ for any d, which should not come as a surprise since Algorithms 2 and 3 perform identically for $p = 1$. Comparing $\mathbb{E}_{\mathtt{mb}}[2, d]$ and $\mathbb{E}_{\mathtt{ds}}[2, d]$, however, we observe that

$$\frac{\sqrt{2}}{\sqrt{\pi}} \approx 0.797 < 0.886 \approx \frac{\sqrt{\pi}}{2},$$

implying that Algorithm 3 is providing a higher expected decrease than Algorithm 2 when a two-dimensional subspace is used.

We end this subsection with asymptotic results akin to Corollary 3.5, that follows from combining Lemma 3.4 with Corollary 4.3.

Corollary 4.4. *Under the same assumptions as Corollary 4.3, asymptotically*

$$\mathbb{E}_{\mathtt{mb}}[1, d_1] \to \frac{\sqrt{2}}{\sqrt{\pi}\sqrt{d_1}} \text{ as } d_1 \to \infty,$$

and

$$\mathbb{E}_{\mathtt{mb}}[2, d_1] \to \frac{\sqrt{\pi}}{\sqrt{2}\sqrt{d_1}} \text{ as } d_1 \to \infty.$$

4.2 Expected decrease per function evaluation

We now examine the expected decrease guarantee of Algorithm 3 by taking its function evaluation cost into account. While Algorithm 2 was evaluating $2p$ new points per iteration, Algorithm 3 only evaluates $p + 1$ new points per iteration. Indeed, the construction of the simplex gradient requires $p + 1$ function values but only p new ones since that of the incumbent solution x^k is re-used from the past iteration. One final evaluation is used in line 6 of Algorithm 3, so the total amounts to $p + 1$ new evaluations. As a result, we define

$$\mathbb{E}^F_{\mathtt{mb}}[p, d] = \frac{\mathbb{E}_{\mathtt{mb}}[p, d]}{p + 1} \tag{4.4}$$

for $p \geq 2$, and investigate its behavior as p varies in Theorem (4.5) (the case $p = 1$ will be discussed separately).

Theorem 4.5. *Under the same assumptions as Theorem 4.2, suppose further than* $d > 2$. *Then, for any* $p = 2, \ldots, d - 1$,

$$\frac{\mathbb{E}_{\mathtt{mb}}[p, d]}{p + 1} > \frac{\mathbb{E}_{\mathtt{mb}}[p + 1, d]}{p + 2}. \tag{4.5}$$

Proof. To obtain the desired result, it suffices to prove that

$$\frac{\mathbb{E}_{\mathtt{mb}}[p,d]}{\mathbb{E}_{\mathtt{mb}}[p+1,d]} = \frac{\Gamma(p/2+1/2)^2}{\Gamma(p/2)\Gamma(p/2+1)} > \frac{p+1}{p+2}.$$

To this aim, we require a tighter version of Gautschi's inequality than the one used to prove Lemma 3.6. By Kershaw's extension to Gautschi's inequality [14], for all $x > 0$ and $s \in (0,1)$, it holds that

$$(x+s/2)^{1-s} < \frac{\Gamma(x+1)}{\Gamma(x+s)} < \left(x-1/2+(s+1/4)^{1/2}\right)^{1-s}. \tag{4.6}$$

Applying $x = p/2$ and $s = 1/2$ in equation (4.6) yields

$$\frac{\Gamma(p/2+1)}{\Gamma(p/2+1/2)} < \frac{\sqrt{p+\sqrt{3}-1}}{\sqrt{2}}.$$

Using $\Gamma(p/2+1) = (p/2)\Gamma(p/2)$, we also have

$$\frac{\Gamma(p/2)}{\Gamma(p/2+1/2)} < \frac{\sqrt{2}\sqrt{p+\sqrt{3}-1}}{p}.$$

Inverting both inequalities and multiplying the results shows that

$$\frac{\Gamma(p/2+1/2)^2}{\Gamma(p/2)\Gamma(p/2+1)} > \frac{p}{p+\sqrt{3}-1}.$$

We can easily verify that $\frac{p}{p+\sqrt{3}-1} \geq \frac{p+1}{p+2}$ whenever $p \geq \sqrt{3}+1 \approx 2.73$. The case of $p = 2$ is easily checked, as

$$\frac{\mathbb{E}_{\mathtt{mb}}[2,d]}{\mathbb{E}_{\mathtt{mb}}[3,d]} = \frac{\pi}{4} > \frac{2+1}{3+1},$$

and therefore (4.5) holds. □

The result of Theorem 4.2 leads to similar conclusions than that of Theorem 3.8, in the sense that using low-dimensional subspace dimension leads to better expected decrease guarantees up to $p \geq 2$. We comment thereafter on other settings.

The case $p = 1$: The inequality (4.5) does not apply for $p = 1$, as

$$\frac{\mathbb{E}_{\mathtt{mb}}[1,d]}{\mathbb{E}_{\mathtt{mb}}[2,d]} = \frac{2}{\pi} < \frac{1+1}{2+1},$$

seemingly indicating that $p = 2$ is the best choice. However, when $p = 1$, the simplex gradient is necessarily equal to b_1 or $-b_1$. In the former case, Algorithm 3 will not require an additional value on line 6, since the value at $z + \delta u = z + \delta b_1$ was already computed and used to form the simplex gradient.

As a result, the average number of function evaluations used when $p = 1$ is $3/2$ (similar to the case of opportunistic polling discussed in Section 3.2). By extending (4.4) to $p = 1$ using this cost, we obtain

$$\mathbb{E}^F_{\texttt{mb}}[1, d] := \frac{\mathbb{E}_{\texttt{mb}}[1, d]}{3/2} = \frac{2}{3\sqrt{\pi}} \frac{\Gamma(d/2)}{\Gamma(d/2 + 1/2)} > \mathbb{E}^F_{\texttt{mb}}[2, d],$$

suggesting that one-dimensional subspaces also provide a better return on investment in model-based approaches based on simplex gradients, i.e., linear models of the function.

Parallel processing: Similarly to the direct-search case, we can consider the situation where c parallel cores are used to compute distinct function evaluations. This paradigm reduces the per-iteration cost of Algorithm 3 to $\lceil p/c \rceil + 1$, where the gain is necessarily achieved only on the evaluations used to form the simplex gradient (the final evaluation on Line 6 must be done after the others). Then, assuming p/c is an integer, we obtain

$$\frac{\mathbb{E}_{\texttt{mb}}[p, d]}{\lceil p/c \rceil + 1} = c\frac{\mathbb{E}_{\texttt{mb}}[p, d]}{p + c} \quad \text{and} \quad \frac{\mathbb{E}_{\texttt{mb}}[p + c, d]}{\lceil (p + c)/c \rceil + 1} = c\frac{\mathbb{E}_{\texttt{mb}}[p + c, d]}{p + 1 + c}.$$

Although the result of Theorem 4.5 does not directly apply to this new quantity (unless $c = 1$), a simple numerical inspection confirms that $\frac{\mathbb{E}_{\texttt{mb}}[p,d]}{\lceil p/c \rceil + 1}$ is maximized for $p = c$ for all values $c \in \{1, 2, \ldots, 256\}$ and $p \in \{c, 2c, \ldots, 100c\}$. This strongly suggests that the expected decrease per unit of work is maximized when you use the smallest subspace that uses all cores, as in the direct-search setting. However, this maximum is not uniquely obtained, since when $d \geq 4$ and $c = 2$, we have

$$\frac{\mathbb{E}_{\texttt{mb}}[2, d]}{\lceil 2/2 \rceil + 1} = \frac{\sqrt{\pi}}{4} \cdot \frac{\Gamma(d/2)}{\Gamma(d/2 + 1/2)} = \frac{\mathbb{E}_{\texttt{mb}}[4, d]}{\lceil 4/2 \rceil + 1},$$

hence both $p = 2$ and $p = 4$ achieve the maximum expected decrease per unit of work.

5 Numerical estimation of expected decrease

In Sections 3 and 4, we showed that the expected decrease per function evaluation is strictly decreasing as a function of p. Considering Corollaries 3.3 and 4.3, we see that the expected decrease improves from $p = 1$ to $p = 2$. Indeed,

$$\frac{\mathbb{E}_{\texttt{mb}}[2, d]}{\mathbb{E}_{\texttt{mb}}[1, d]} = \frac{\pi}{2} > 1 \quad \text{and} \quad \frac{\mathbb{E}_{\texttt{ds}}[2, d]}{\mathbb{E}_{\texttt{ds}}[1, d]} = \sqrt{2} > 1.$$

However, the expected decrease per function evaluation actually worsens from $p = 1$ to $p = 2$. Indeed,

$$\frac{\mathbb{E}^F_{\texttt{ds}}[2, d]}{\mathbb{E}^F_{\texttt{ds}}[1, d]} = \sqrt{2}/2 < 1 \quad \text{and} \quad \frac{\mathbb{E}^F_{\texttt{mb}}[2, d]}{\mathbb{E}^F_{\texttt{mb}}[1, d]} = \frac{\pi}{4} < 1.$$

Computing these ratios becomes increasingly cumbersome as p increases. In this section, we thus investigate the behavior of the expected decrease quantities $\mathbb{E}_{\texttt{ds}}$, $\mathbb{E}_{\texttt{mb}}$, $\mathbb{E}^F_{\texttt{ds}}$ and $\mathbb{E}^F_{\texttt{mb}}$ numerically, by way of to Monte Carlo simulations.

Algorithm 4 describes our estimation procedure applied to evaluate the expected decrease quantities. Note that it samples both a vector g uniformly distributed on the unit sphere and a random basis B for the subspace, as in the original definition (2.6). (As such, we also numerically verify the results in Proposition 3.1 and 4.1.) The estimated quantity is obtained by averaging the decrease formulas for every sample (g, B). In our subsequent experiments, we use $N_{\texttt{sims}} = 10^4$ samples.

Algorithm 4 Monte Carlo estimation of expected decrease `MCestim`)

procedure MCTEST($N_{\texttt{sims}}$, b, p, `DFi`)
  % $N_{\texttt{sims}}$ number of simulations to run, positive integer
  % d problem dimension, positive integer
  % p subspace dimension, $p \in \{1, 2, \ldots, d\}$
  % `DFi`: DFO step on subspaces `DFi` $\in \{$`ds`,`mb`$\}$
  for $k = 1$ to $N_{\texttt{sims}}$ **do**
    Randomly select $g \in \mathcal{S}^{d-1}$
    Randomly select a subspace of dimension p with orthonormal basis

$$B = [b_1, b_2, ..., b_p]$$

    if `DFi`=`ds` **then**
      Set $D(k) = \max\{g^{\mathrm{T}} d : d = \pm b_i, i = 1, 2, \ldots, p\}$
    else
      Compute the subspace gradient $\hat{g} = B^{\mathrm{T}} g$
      Set $D(k) = (-B(\hat{g}/\|\hat{g}\|))^{\mathrm{T}} g$
    end if
  end for
  Return $\sum_{k=1}^{N_{\texttt{sims}}} D(k)/N_{\texttt{sims}}$ as an estimate of $\mathbb{E}_{\texttt{DFi}}[p, d]$
end procedure

5.1 Direct-search case

We first look at the results for estimating $\mathbb{E}_{\texttt{ds}}$. Note that we can compute the integral symbolically for low values of p using Mathematica [13], yielding

$$\begin{aligned}
\mathbb{E}_{\texttt{ds}}[3, d] &= \frac{\Gamma(d/2)}{\Gamma(d/2+1/2)} \left[\frac{12 \arctan(\sqrt{2}) + 3 \arctan(460\sqrt{2}/329)}{2\sqrt{2}(\sqrt{\pi})^3} \right] &&\approx 0.938 \frac{\Gamma(d/2)}{\Gamma(d/2+1/2)}, \\
\mathbb{E}_{\texttt{ds}}[4, d] &= \frac{\Gamma(d/2)}{\Gamma(d/2+1/2)} \left[\frac{12\sqrt{2} \arctan(\frac{1}{2\sqrt{2}})}{(\sqrt{\pi})^3} \right] &&\approx 1.036 \frac{\Gamma(d/2)}{\Gamma(d/2+1/2)}.
\end{aligned}$$

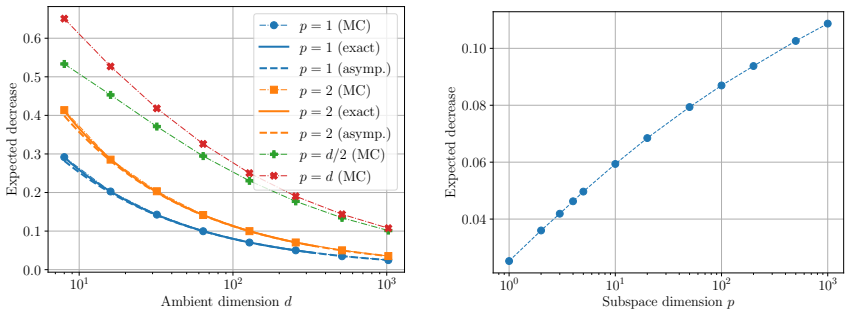
Further estimation of the Gamma functions leads to the approximations

$$\frac{\mathbb{E}^F_{\texttt{ds}}[3, d]}{\mathbb{E}^F_{\texttt{ds}}[2, d]} \approx 0.784 \quad \text{and} \quad \frac{\mathbb{E}^F_{\texttt{ds}}[4, d]}{\mathbb{E}^F_{\texttt{ds}}[3, d]} \approx 0.828.$$

These values suggest that the gain in expected decrease between p and $p+1$ reduces as the value of p increases.

Numerical estimations of the expected decrease for direct-search are given in Figure 1. [1] Figure 1a presents the output of Algorithm 4 (with `DFi=ds`) for varying dimensions $d \in \{8, 16, 32, \ldots, 1024\}$ using subspace dimension $p \in \{1, 2, d/2, d\}$. For $p \in \{1, 2\}$ we superimpose the exact formula for the expected decrease as given by Corollary 3.3. For large values of d, floating-point and overflow errors occur when evaluating $\Gamma(d/2)/\Gamma(d/2+1/2)$, thus we only plot the values from Corollary 3.3 up to occurrence of these errors. For comparison, we also show the large-d asymptotic results from Corollary 3.5. We note that the Monte-Carlo simulation aligns nearly perfectly with the formulas for $p \in \{1, 2\}$, while the large-d asymptotics are essentially indistinguishable from the simulations for $d \geq 100$.

Figure 1b shows the output of Algorithm 4 (with `DFi=ds`) for varying subspace size $p \in \{1, 2, 3, 4, 5, 10, 20, 50, 100, 200, 500, 1000\}$ and fixed dimension $d = 1000$. As expected, we observe that choosing $p = 1$ provides the worst expected decrease and that $p = d$ leads to the best expected decrease. Note also that the expected decrease diminishes as d increases.



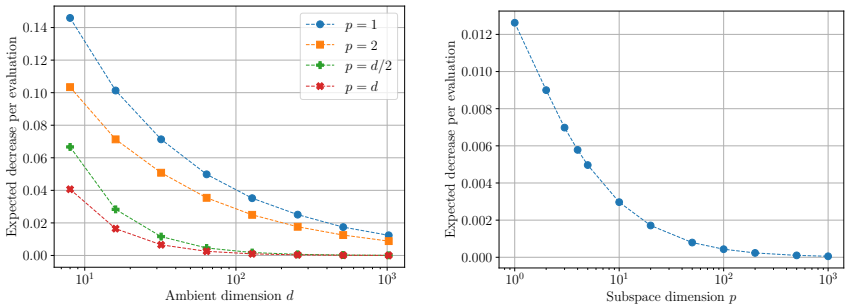
(a) Varying dimension d (b) Varying subspace dimension p

Figure 1: Expected decrease ($\mathbb{E}_{\texttt{ds}}[p, d]$) versus average decrease based on Monte Carlo simulation for varying dimension (a) and subspace dimension (b).
(a) Lines with "(MC)" are the Monte Carlo simulation results, "(exact)" is the result from Theorem 3.2 and "(asymp.)" is the large-d asymptotic result from Corollary 3.5.
(b) Ambient dimension $d = 1000$.

In Theorem 3.8, we showed that the expected decrease per function evaluation $\mathbb{E}^F_{\texttt{ds}}[p, d]$ was strictly decreasing as a function of the subspace dimension p. In Figure 2, we plot the expected decrease per unit work for varying dimensions and varying subspace dimensions. Those results confirm our theoretical findings, in that setting $p = 1$ gives the largest expected decrease per function evaluation. Note that the gap between $\mathbb{E}^F_{\texttt{ds}}[p, d]$ and $\mathbb{E}^F_{\texttt{ds}}[p, d]$ is the largest for

[1] In all figures it should be recognized that lines adjoining points are for visualization only. The values of d and p are always integers.

$p = 1$, and that it decreases as p increases.

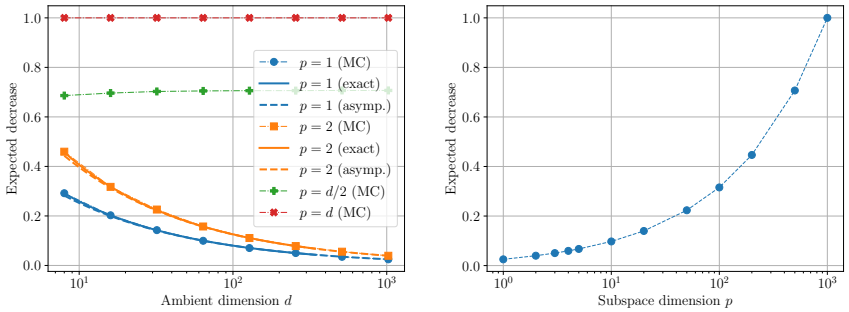


(a) Varying dimension d (b) Varying subspace dimension p

Figure 2: Expected decrease per function evaluation $\mathbb{E}^F_{\texttt{ds}}[p, d]$ (3.2) versus average decreased based on Monte Carlo simulation for varying dimension (a) and subspace dimension (b).

5.2 Model-based case

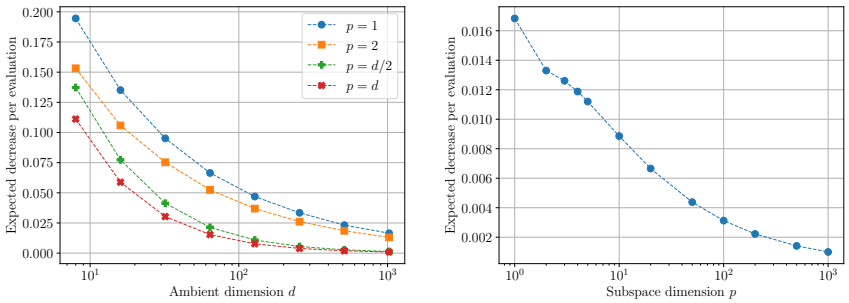
We now discuss the output of Algorithm 4 using `DFi=mb`. Figures 3 and 4 present results analogous to that of Figures 1 and 2.



(a) Varying dimension d (b) Varying subspace dimension p

Figure 3: Expected decrease ($\mathbb{E}_{\texttt{mb}}[p, d]$) versus the average decreased based on Monte Carlo simulation for varying dimension (a) and subspace dimension (b).

As in the direct-search case, we match exact results for $p \in \{1, 2\}$ (see Corollary 4.3) and large d-asymptotics (see Corollary 4.4) quite closely. We also observe empirically that $p = 1$ is worst in terms of expected decrease but best in terms of expected decrease per function evaluation (with our choice of $\mathbb{E}^F_{\texttt{mb}}[1, d] = \mathbb{E}_{\texttt{mb}}[1, d]/(3/2)$ explained in Section 4.2). Finally, we see from Figure 4b that the gap between $p = 1$ and $p = 2$ is the largest among all consecutive values of p.



(a) Varying dimension d (b) Varying subspace dimension p

Figure 4: Expected decrease per unit work $\mathbb{E}^F_{\mathtt{mb}}[p, d]$ versus the average based on Monte Carlo simulation for varying dimension (a) and subspace dimension (b).

6 Discussion

We have established expected decrease formulae for derivative-free iterations using random subspaces when applied to linear functions. As explained in Section 2.3, our analysis can be employed to show expected decrease guarantees for more general classes of smooth functions that admit a linear model approximation. We have established that performing iterations of derivative-free algorithms in randomly generated subspaces is more beneficial as the dimension of the subspaces decreases. This arguably surprising result arises from properties of the uniform distribution over subspaces, and goes some way to understanding the strong empirical performance of low-dimensional subspace approximations (e.g. in [12, 22]).

Extending our analysis to handle quadratic models is a natural continuation of this paper, that poses a number of challenges related to the theory of random quadratic functions. Nevertheless, such results seem necessary to understand derivative-free methods that rely on quadratic models and beyond. In addition, elaborate implementations of derivative-free algorithms can reuse past evaluations to produce better trial points, which introduces non-trivial dependencies between iterations. Finally, we expect our theory to apply in the case of stochastic function evaluations, provided those satisfy common probabilistic properties appearing in the literature.

References

[1] C. Audet and W. Hare. *Derivative-Free and Blackbox Optimization*. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, 2017.

[2] E. Bergou, E. Gorbunov, and P. Richtárik. Stochastic three points method for unconstrained smooth minimization. *SIAM J. Optim.*, 30:2726–2749,

2020.

[3] C. Cartis, J. Fowkes, and Z. Shao. Randomised subspace methods for non-convex optimization, with applications to nonlinear least-squares. arXiv::2211.09873, 2022.

[4] C. Cartis and L. Roberts. Scalable subspace methods for derivative-free nonlinear least-squares optimization. *Math. Program.*, 199:461–524, 2023.

[5] Y. Chikuse. *Statistics on Special Manifolds*. Lecture Notes in Statistics. Springer, New York, 2003.

[6] A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to Derivative-Free Optimization*. MPS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, 2009.

[7] M. A. Diniz-Ehrhardt, J. M. Martínez, and M. Raydan. A derivative-free nonmonotone line-search technique for unconstrained optimization. *J. Comput. Appl. Math.*, 219:383–397, 2008.

[8] *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.1.8 of 2022-12-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[9] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: the power of two function evaluations. *IEEE Trans. Inform. Theory*, 61:2788–2806, 2015.

[10] K. J. Dzahini and S. M. Wild. Stochastic trust-region algorithm in random subspaces with convergence and expected complexity analyses. arXiv:2207.06452, 2022.

[11] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20:303–353, 1998.

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

[13] Wolfram Research, Inc. Mathematica, Version 13.2. Champaign, IL, 2022.

[14] D. Kershaw. Some extensions of W. Gautschi's inequalities for the gamma function. *Mathematics of Computation*, 41:607–611, 1983.

[15] T. G. Kolda, R. M. Lewis, and V. Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Rev.*, 45:385–482, 2003.

[16] D. Kozak, S. Becker, A. Doostan, and L. Tenorio. A stochastic subspace approach to gradient-free optimization in high dimensions. *Comput. Optim. Appl.*, 79:339–368, 2021.

[17] D. Kozak, C. Molinari, L. Rosasco, L. Tenorio, and S. Villa. Zeroth-order optimization with orthogonal random directions. *Math. Program.*, 199:1179–1219, 2023.

[18] J. Larson, M. Menickelly, and S. M. Wild. Derivative-free optimization methods. *Acta Numer.*, 28:287–404, 2019.

[19] M. Menickelly. Avoiding geometry improvement in derivative-free model-based methods via randomization. arXiv:2305.17336, 2023.

[20] V. D. Milman and G. Schechtman. *Asymptotic Theory of Finite Dimensional Normed Spaces*. Lecture Notes in Mathematics. Springer Berlin, Heidelberg, 1986.

[21] Yu. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17:527–566, 2017.

[22] L. Roberts and C. W. Royer. Direct search based on probabilistic descent in reduced subspaces. *SIAM J. Optim.*, 2023 (To appear).

[23] Z. Shao. *On Random Embeddings and their Applications to Optimization*. PhD thesis, University of Oxford, 2022.