UCLouvain

École polytechnique de Louvain

Random embeddings for global optimization: Convergence results beyond isotropy

Author: **Roy MAKHLOUF**
Supervisor: **Estelle MASSART**
Readers: **Geovani GRAPIGLIA, Laurent JACQUES, Estelle MASSART**
Academic year 2024–2025
Master [120] in Mathematical Engineering

Abstract

Many real-world optimization problems are high-dimensional, requiring dimensionality reduction techniques to solve them efficiently. Recently, the use of random embeddings was shown to substantially outperform classical methods for Lipschitz continuous objectives with special structure, such as functions with low effective dimension. Tools from conic integral geometry have been used to explore the benefits of random embeddings for global optimization of Lipschitz continuous objectives with no additional structure. These tools allow to derive lower bounds on the probability that a random linear subspace intersects a ball of approximate minimizers, by using the circular cone tangent to the ball. We aim here to extend these results to functions that vary very slowly along a linear subspace, for which we replace the ball of approximate minimizers with an ellipsoid to account for the anisotropic structure of the objective. Our findings offer deeper insights into how the anisotropic structure in high-dimensional functions impacts optimization algorithms.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Estelle Massart, for her availability throughout the year, her guidance and her valuable insights. Her expertise and thoughtful feedbacks were of great help in shaping the direction and quality of this thesis.

I would also like to thank the readers of this thesis, Professor Geovani Grapiglia and Professor Laurent Jacques.

Lastly, many thanks to my family and friends for their support and presence during the year. Their belief in me has been a constant source of motivation throughout this beautiful journey.

Contents

List of Figures

Chapter 1

Introduction

Optimization in high-dimensional spaces is inherent to many real-world applications, including machine learning, scientific computing and financial modeling. However, high dimensionality introduces significant challenges, such as the curse of dimensionality, which often renders classical optimization algorithms inefficient for large-scale problems. In our era of big data, dimensionality reduction techniques have become more crucial than ever for solving these problems effectively.

A promising approach that has recently gained attention for dimensionality reduction is the use of random embeddings. These techniques project high-dimensional optimization problems into lower-dimensional linear subspaces where optimization algorithms become more tractable. Random embeddings have been shown to substantially outperform classical methods for Lipschitz continuous objectives with special structure, named functions with low effective dimensionality. These functions, although defined in a high-dimensional space, only vary along a low-dimensional linear subspace. Random embeddings have been successfully applied to overparameterized problems, such as deep neural network training, where it is widely believed that the objective possesses low effective dimensionality due to the high dimensionality of the search space.

To better understand the behavior of optimization algorithms, it is natural to derive associated convergence results, such as the minimum number of iterations required to solve a problem or a lower bound on the probability that a randomized optimization algorithm solves the problem, where the word "solve" could have different meaning depending on the context. For Lipschitz continuous objectives with no additional structure, tools from conic integral geometry have been employed to derive lower

bounds on the probability that a random linear subspace intersects a ball of approximate minimizers, by using the circular cone tangent to the ball. Extending this analysis to functions with low effective dimensionality has shed light on why random embeddings are especially effective for such functions.

In an effort to generalize the previous results, we derive in this thesis convergence results, specifically probability lower bounds, for the random embeddings with Lipschitz continuous functions that vary very slowly along a linear subspace, which we refer to as functions with anisotropic dimensionality or AD functions. AD functions vary anisotropically, meaning they may change rapidly in some directions and much more slowly in others. In the following work, we extend existing results for functions with low effective dimensionality to AD functions by replacing the isotropic model of approximate minimizers (i.e. the ball) with an ellipsoid, in order to capture the anisotropic structure of the objective.

1.1 Problem definition

We consider the global optimization problem

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

where $f : \mathcal{X} \to \mathbb{R}$ is Lipschitz continuous (with possibly some additional structure) and $\mathcal{X} \subseteq \mathbb{R}^D$ is a domain with non-empty interior. Here, $D \in \mathbb{N}$ can be very large, implying that we are dealing with a large-scale optimization problem.

We recall the definition of a Lipschitz continuous function:

Definition 1.1.1 A function $f : \mathcal{X} \to \mathbb{R}$ is called Lipschitz continuous if there exists a real constant $L \geq 0$ such that

$$|f(x) - f(y)| \leq L\|x - y\|$$

for all $x, y \in \mathcal{X}$.

Lipschitz continuity is a natural and common assumption in optimization, as it allows us to have some information over how the objective function behaves with respect to its inputs and enables us to derive theoretical guarantees for optimization algorithms.

Note that very few assumptions have been made concerning the domain $\mathcal{X}$, which may be non-convex or even equal to the entire space $\mathbb{R}^D$. In this thesis, we focus on problems with large values of D, as we want to solve large-scale optimization tasks that frequently arise in various fields such as machine learning, engineering and natural sciences.

1.2 State of the art

As a means of reducing the dimensionality of the search space, this thesis primarily focuses on random embeddings, a class of transformations that map high-dimensional spaces into lower-dimensional ones using randomness. More precisely, rather than solving the original global optimization problem in $\mathbb{R}^D$, we instead consider (possibly a sequence of) randomized reduced problems:

$$f^*_{A,p} := \min_{y \in \mathbb{R}^d} f(Ay + p)$$

$$\text{s.t.} \quad Ay + p \in \mathcal{X}$$

where $A \in \mathbb{R}^{D\times d}$ is a Gaussian matrix, $d \ll D$ and $p \in \mathbb{R}^D$ is arbitrary or user-defined. A visual illustration of the concept of a random embedding is shown in Figure 1.1.

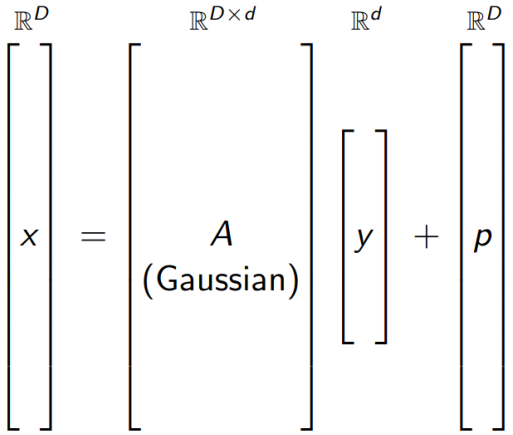


Figure 1.1: Illustration of the concept of random embeddings

We are interested in deriving lower bounds on the probability that solving the randomized reduced problem yields an ε-minimizer of the original global optimization problem. That is, we seek a feasible point

$$x \in (p + \text{range}(A)) \cap \mathcal{X}$$

satisfying

$$f(x) \leq f^\star + \varepsilon$$

for some prescribed accuracy $\varepsilon > 0$. Under the additional assumption that there exists of a D-dimensional closed Euclidean ball of radius ε/L centered at a global minimizer x^* such that

$$B_{\varepsilon/L}[x^*] \subset \mathcal{X}$$

and using tools from conic integral geometry, the authors of [10] have been able to derive the bound

$$\mathbb{P}[f^*_{A,p} \leq f^* + \varepsilon] \geq \tau(r_p, d, D), \tag{1.1}$$

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

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

and with $r_p := \varepsilon/(L\|x^* - p\|)$ as well as $p \in \mathcal{X} \setminus \{x \in \mathcal{X} : f(x) \leq f^* + \varepsilon\}$.

In addition, by analyzing the asymptotic behavior of $\tau(r, d, D)$, the authors obtained an asymptotic expansion of the lower bound. In particular, they showed that

$$\tau(r_p, d, D) = \Theta\left(D^{\frac{d-2}{2}}\left(\frac{\varepsilon}{L\|x^* - p\|}\right)^{D-d}\right)$$

as $D \to \infty$.

1.3 Contributions

Unless explicitly stated otherwise, all results presented in this work, including statements of theorems, propositions, proofs, etc are derived by the author. If a result is not original to the author, a reference to the source article is generally provided.

Chapter 2

Theoretical background

2.1 Convex geometry

Throughout this thesis, we will often draw upon tools from convex geometry, including equations and representations of common geometric objects. In particular, our focus will be on convex shapes, which are essential for the following chapters.

Definition 2.1.1 A set $C \subseteq \mathbb{R}^n$ is called convex if the line segment between any two points of C lies in C i.e. for all $x, y \in C$ and for all $\lambda \in [0, 1]$,

$$\lambda x + (1 - \lambda)y \in C.$$

2.1.1 Ball

This is one of the most known geometric shape in Euclidean spaces. The closed n-dimensional Euclidean ball of radius $r \in \mathbb{R}_0^+$ and center $c \in \mathbb{R}^n$ is given by

$$B_r(c) := \{x \in \mathbb{R}^n \mid \|x - c\| \leq r\}.$$

It is the set of all points in $\mathbb{R}^n$ that are at a distance less than r from c.

2.1.2 Ellipsoid

An ellipsoid is a geometric object that can be obtained by applying an affine transformation to a ball. Formally, a standard ellipsoid is defined as

$$E_\Lambda = \{x \in \mathbb{R}^n \mid x^\top \Lambda x \leq 1\},$$

where $\Lambda \in \mathbb{R}^{n \times n}$ is a positive definite diagonal matrix. The diagonal entries of Λ are the reciprocals of the squares of the semi-axis lengths of the ellipsoid. This

set of points represents an ellipsoid centered at the origin, with its principal axes aligned with the standard coordinate axis of the Euclidean space $\mathbb{R}^n$. However, we are interested in more general ellipsoids that may be both translated and rotated in space. A general form of such an ellipsoid, centered at a point $c \in \mathbb{R}^n$, is given by

$$\mathcal{E}_A(c) = \{x \in \mathbb{R}^n \mid (x-c)^\top A(x-c) \leq 1\},$$

where $A = Q\Lambda Q^\top$, with $Q \in \mathbb{R}^{n\times n}$ being an orthogonal matrix whose columns are the eigenvectors of A, corresponding to the principal axis of the ellipsoid. Note that replacing c by the zero vector and Q by the identity matrix gives back the standard ellipsoid E_Λ. Theorem 2.1.1 below shows that an ellipsoid is indeed an affine transformation of an Euclidean ball.

Theorem 2.1.1 Let $B_r(b) \subset \mathbb{R}^n$ be an Euclidean ball centered at $b \in \mathbb{R}^n$ with radius $r \in \mathbb{R}_0^+$ and $\mathcal{E}_A(c) \subset \mathbb{R}^n$ be an ellipsoid centered at $c \in \mathbb{R}^n$, where $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix with eigendecomposition $A = Q\Lambda Q^\top$. It holds

$$\mathcal{E}_A(c) = MB_r(b) + p$$

where

$$M := r^{-1}Q\Lambda^{-1/2} \quad \text{and} \quad p := c - Mb$$

Proof. Let $x \in \mathbb{R}^n$ and $y = Mx + p$. There holds

$$\begin{aligned}(y-c)^\top A(y-c) &= (r^{-1}Q\Lambda^{-1/2}(x-b))^\top Q\Lambda Q^\top (r^{-1}Q\Lambda^{-1/2}(x-b)) \\ &= r^{-2}(x-b)^\top \Lambda^{-1/2}Q^\top Q\Lambda Q^\top Q\Lambda^{-1/2}(x-b) \\ &= \frac{\|x-b\|^2}{r^2}.\end{aligned}$$

Therefore, if $x \in B_r(b)$, by definition of an Euclidean ball, we have

$$\|x-b\| \leq r$$

and so

$$(y-c)^\top A(y-c) \leq 1,$$

which means $y \in \mathcal{E}_A(c)$. Similarly, if $y \in \mathcal{E}_A(c)$, then

$$(y-c)^\top A(y-c) \leq 1$$

and so

$$\|x-b\| \leq r,$$

which means $x \in B_r(b)$. Hence, it follows that

$$\mathcal{E}_A(c) = MB_r(b) + p$$

and so an ellipsoid is indeed an affine transformation of an Euclidean ball. □

2.1.3 Circular cone

Besides the bounded convex geometric shapes that we saw earlier (usually called convex bodies), we now look at unbounded geometric shapes called cones. For this, the following definition will be useful.

Definition 2.1.2 A set $C \subseteq \mathbb{R}^n$ is called a cone if $x \in C$ implies $\lambda x \in C$ for all $\lambda \geq 0$.

More precisely, we will consider cones that are convex.

Proposition 2.1.1 A set C is a convex cone if and only if for all $x, y \in C$ and for all $\alpha, \beta \geq 0$,

$$\alpha x + \beta y \in C.$$

The proposition above easily follows by combining the definition of a convex set and a cone.

The first type of convex cone that we will look at is called the circular cone with angle $\alpha \in [0, \frac{\pi}{2}]$ and is defined as

$$\mathrm{Circ}_n(\alpha) := \{x \in \mathbb{R}^n \mid x_1 \geq \|x\| \cos(\alpha)\}.$$

The circular cone that we consider is the one that has the basis vector e_1 as principal axis. We could also consider the one that has an arbitrary unit vector $v \in \mathbb{R}^n$ as principal axis and its description would be

$$\{x \in \mathbb{R}^n \mid \langle x, v\rangle \geq \|x\| \cos(\alpha)\}$$

but we chose to deal with the first representation instead. Note that the circular cone with angle α is just the set of vectors that form an angle of at most α with e_1. In fact, by the geometric definition of the dot product, we know that for all $x, y \in \mathbb{R}^n$,

$$\langle x, y\rangle = \|x\|\|y\| \cos(\alpha)$$

where α is the angle between x and y. Let $x \in \mathbb{R}^n_0$ be an arbitrary vector that forms an angle of $\beta \in [0, \frac{\pi}{2}]$ with e_1. Then there holds

$$\langle x, e_1\rangle = \|x\|\|e_1\| \cos(\beta) = \|x\| \cos(\beta)$$

and so

$$\cos(\beta) = \frac{\langle x, e_1\rangle}{\|x\|}.$$

Requiring $\beta \leq \alpha$ is equivalent to require $\cos(\beta) \geq \cos(\alpha)$ as $\cos(\cdot)$ is a decreasing function on $[0, \frac{\pi}{2}]$. Hence, it follows that

$$\frac{\langle x, e_1 \rangle}{\|x\|} \geq \cos(\alpha)$$

and so

$$x_1 \geq \|x\| \cos(\alpha)$$

which explains the rationale behind the given representation of the circular cone.

2.1.4 Elliptical cone

For our purposes, we simply define an elliptical cone to be a linear transformation of a circular cone i.e.

$$\mathcal{A}\ \mathrm{Circ}_n(\alpha)$$

with $\mathcal{A} \in \mathbb{R}^{n \times n}$ a positive definite matrix and $\alpha \in [0, \frac{\pi}{2}]$. Elliptical cones play a central role in this thesis, as we will use them several times in our analysis. Specifically, we will often require an equation of the elliptical cone tangent to a given ellipsoid in the Euclidean n-space, with the apex locate at the origin. This is the focus of Lemma 2.1.1.

Lemma 2.1.1 Let

$$\mathcal{E}_A(c) := \{x \in \mathbb{R}^n \mid (x - c)^\top A(x - c) \leq 1\}$$

be an ellipsoid that doesn't contain the origin. Then the elliptical cone emanating from the origin and tangent to the ellipsoid is given by

$$\mathcal{K}_{\mathcal{E}_A(c)} := \{x \in \mathbb{R}^n \mid (x^\top Ac)^2 \geq (x^\top Ax)(c^\top Ac - 1),\ x^\top Ac \geq 0\}.$$

Proof. We consider rays λx with $\lambda \in \mathbb{R}^+$ and $x \in \mathbb{R}^n$ emanating from the origin that intersect the ellipsoid. In mathematical terms, this means that the equation

$$(\lambda x - c)^\top A(\lambda x - c) = 1$$

must have one or two solutions λ for each x in the elliptical cone as each ray in such cone can either be tangent to the ellipsoid or intersect its boundary twice. Note that

$$(\lambda x - c)^\top A(\lambda x - c) = 1$$

$$\Leftrightarrow (\lambda x - c)^\top (\lambda Ax - Ac) = 1$$

$$\Leftrightarrow (x^\top Ax)\lambda^2 - (x^\top Ac)\lambda - (c^\top Ax)\lambda + c^\top Ac = 1$$
$$\Leftrightarrow (x^\top Ax)\lambda^2 - 2(x^\top Ac)\lambda + (c^\top Ac - 1) = 0$$

which is a quadratic equation in terms of λ. In order to have one or two solutions to this equation, we need a nonnegative discriminant i.e.

$$\Delta = 4(x^\top Ac)^2 - 4(x^\top Ax)(c^\top Ac - 1) \geq 0. \tag{2.1}$$

The solution(s) of the quadratic equation can be expressed as

$$\lambda_1 = \frac{2(x^\top Ac) - \sqrt{\Delta}}{2(x^\top Ax)} \quad \text{and} \quad \lambda_2 = \frac{2(x^\top Ac) + \sqrt{\Delta}}{2(x^\top Ax)}.$$

Since A is positive definite, there holds

$$x^\top Ax > 0$$

and so for λ_1 to be nonnegative, we must have

$$x^\top Ac \geq 0. \tag{2.2}$$

With condition (2.2), λ_2 is always nonnegative so let's focus on λ_1 to see if more conditions are needed. Since $0 \notin \mathcal{E}_A(c)$, we have

$$(0 - c)^\top A(0 - c) = c^\top Ac > 1$$

and so

$$\Delta \leq 4(x^\top Ac)^2.$$

This in turn implies that

$$2(x^\top Ac) \geq \sqrt{\Delta}$$

and hence, λ_1 is nonnegative as well under the conditions (2.1) and (2.2). The conclusion follows. □

Lemma 2.1.1 is certainly useful but it does not really show how the obtained elliptical cone is a linear transformation of a circular cone. This is captured in the following important theorem.

Theorem 2.1.2 Let

$$\mathcal{E}_A(c) := \{x \in \mathbb{R}^n \mid (x - c)^\top A(x - c) \leq 1\}$$

be an ellipsoid that doesn't contain the origin, where $A \in \mathbb{R}^{n\times n}$ is a positive definite matrix with eigendecomposition $A = Q\Lambda Q^\top$. Then, the elliptical cone emanating from the origin and tangent to the ellipsoid is given by

$$\mathcal{C}_{\mathcal{E}_A(c)} := \mathcal{A}\ \mathrm{Circ}_n\left(\arcsin\left(\frac{1}{\|\Lambda^{1/2}Q^\top c\|}\right)\right),$$

where

$$\mathcal{A} := Q\Lambda^{-1/2}H, \quad H := I - 2\frac{(e_1-u)(e_1-u)^\top}{(e_1-u)^\top(e_1-u)} \quad \text{and} \quad u := \frac{\Lambda^{1/2}Q^\top c}{\|\Lambda^{1/2}Q^\top c\|}.$$

Proof. Let $y \in \mathcal{C}_{\mathcal{E}_A(c)}$. We show that $y \in \mathcal{K}_{\mathcal{E}_A(c)}$. Since $y \in \mathcal{C}_{\mathcal{E}_A(c)}$, this means that there exists

$$x \in \mathrm{Circ}_n\left(\arcsin\left(\frac{1}{\|\Lambda^{1/2}Q^\top c\|}\right)\right)$$

such that

$$y = \mathcal{A}x.$$

There holds

$$\begin{aligned}
(y^\top Ac)^2 &= \left((Q\Lambda^{-1/2}Hx)^\top(Q\Lambda Q^\top)c\right)^2 \\
&= \left(x^\top H^\top\Lambda^{-1/2}Q^\top Q\Lambda Q^\top c\right)^2 \\
&= \left(x^\top H^\top\Lambda^{1/2}Q^\top c\right)^2 \\
&= \left(x^\top H^\top\frac{\Lambda^{1/2}Q^\top c}{\|\Lambda^{1/2}Q^\top c\|}\right)^2\|\Lambda^{1/2}Q^\top c\|^2 \\
&= \left(x^\top H^\top u\right)^2(\Lambda^{1/2}Q^\top c)^\top(\Lambda^{1/2}Q^\top c) \\
&= \left(x^\top H^\top u\right)^2(c^\top Q\Lambda Q^\top c) \\
&= \left(x^\top H^\top u\right)^2(c^\top Ac).
\end{aligned}$$

Note that H is an orthogonal matrix that sends the basis vector e_1 to u (it is a Householder matrix). In fact,

$$\begin{aligned}
He_1 &= \left(I - 2\frac{(e_1-u)(e_1-u)^\top}{(e_1-u)^\top(e_1-u)}\right)e_1 \\
&= e_1 - 2\frac{(e_1-u)(e_1-u)^\top e_1}{(e_1-u)^\top(e_1-u)} \\
&= e_1 - 2\frac{(e_1-u)(1-u_1)}{\|e_1\|^2 - 2u_1 + \|u\|^2} \\
&= e_1 - 2(1-u_1)\frac{e_1-u}{2(1-u_1)} \\
&= e_1 - (e_1-u) \\
&= u.
\end{aligned}$$

Hence,

$$H^\top u = e_1$$

and so

$$(y^\top Ac)^2 = (x^\top e_1)^2 (c^\top Ac) = x_1^2 (c^\top Ac).$$

Note that

$$\begin{aligned} y^\top Ac &= (Q\Lambda^{-1/2} Hx)^\top (Q\Lambda Q^\top) c \\ &= x^\top H^\top \Lambda^{-1/2} Q^\top Q \Lambda Q^\top c \\ &= x^\top H^\top \Lambda^{1/2} Q^\top c \\ &= x^\top H^\top u \cdot \|\Lambda^{1/2} Q^\top c\| \\ &= x^\top e_1 \cdot \|\Lambda^{1/2} Q^\top c\| \\ &= x_1 \cdot \|\Lambda^{1/2} Q^\top c\| \\ &\geq 0 \end{aligned}$$

as x belongs to a circular cone with the basis vector e_1 as principal axis (and so $x_1 \geq 0$). Furthermore, since

$$x \in \mathrm{Circ}_n \left(\arcsin \left(\frac{1}{\|\Lambda^{1/2} Q^\top c\|} \right) \right),$$

there holds

$$\begin{aligned} (y^\top Ac)^2 &= x_1^2 (c^\top Ac) \\ &\geq \|x\|^2 \cos^2 \left(\arcsin \left(\frac{1}{\|\Lambda^{1/2} Q^\top c\|} \right) \right) (c^\top Ac) \\ &= \|x\|^2 \left(1 - \frac{1}{\|\Lambda^{1/2} Q^\top c\|^2} \right) (c^\top Ac) \\ &= \|H^\top \Lambda^{1/2} Q^\top y\|^2 \left(1 - \frac{1}{(\Lambda^{1/2} Q^\top c)^\top (\Lambda^{1/2} Q^\top c)} \right) (c^\top Ac) \\ &= (H^\top \Lambda^{1/2} Q^\top y)^\top (H^\top \Lambda^{1/2} Q^\top y) \left(1 - \frac{1}{c^\top Q \Lambda Q^\top c} \right) (c^\top Ac) \\ &= (y^\top Q \Lambda Q^\top y) \left(1 - \frac{1}{c^\top Ac} \right) (c^\top Ac) \\ &= (y^\top Ay)(c^\top Ac - 1). \end{aligned}$$

This shows that $y \in \mathcal{K}_{\mathcal{E}_A(c)}$ and so $\mathcal{C}_{\mathcal{E}_A(c)} \subseteq \mathcal{K}_{\mathcal{E}_A(c)}$.

Let now $y \in \mathcal{K}_{\mathcal{E}_A(c)}$ and define

$$x := H^\top \Lambda^{1/2} Q^\top y.$$

As

$$(y^\top Ac)^2 \geq (y^\top Ay)(c^\top Ac - 1) \quad \text{and} \quad y^\top Ac \geq 0,$$

we have

$$y^\top Ac \geq \sqrt{(y^\top Ay)(c^\top Ac - 1)}$$

and in a similar reasoning as before, we get

$$x_1\sqrt{c^\top Ac} \geq \|x\| \cos\left(\arcsin\left(\frac{1}{\|\Lambda^{1/2}Q^\top c\|}\right)\right)\sqrt{c^\top Ac}$$

$$\Leftrightarrow x_1 \geq \|x\| \cos\left(\arcsin\left(\frac{1}{\|\Lambda^{1/2}Q^\top c\|}\right)\right)$$

which means that

$$x \in \mathrm{Circ}_n\left(\arcsin\left(\frac{1}{\|\Lambda^{1/2}Q^\top c\|}\right)\right)$$

and so $y \in \mathcal{C}_{\mathcal{E}_A(c)}$. Hence, $\mathcal{K}_{\mathcal{E}_A(c)} \subseteq \mathcal{C}_{\mathcal{E}_A(c)}$ and finally $\mathcal{C}_{\mathcal{E}_A(c)} = \mathcal{K}_{\mathcal{E}_A(c)}$ as expected. □

Remark When $u = e_1$ in Theorem 2.1.2, it is understood that $H = I$.

2.2 Dual cones

Dual cones are a concept often encountered in convex geometry and optimization. Given a set $P \subseteq \mathbb{R}^n$, the dual cone of P is the set of all vectors in $\mathbb{R}^n$ that form an angle of at most $\pi/2$ with every vector in P. Mathematically, this is expressed as follows.

Definition 2.2.1 Let $P \subseteq \mathbb{R}^n$. The dual cone of P, denoted by P^*, is defined as

$$P^* := \{y \in \mathbb{R}^n \mid \langle y, x\rangle \geq 0, \forall x \in P\}.$$

Figure 2.1 offers a geometric interpretation of the concept of a dual cone.

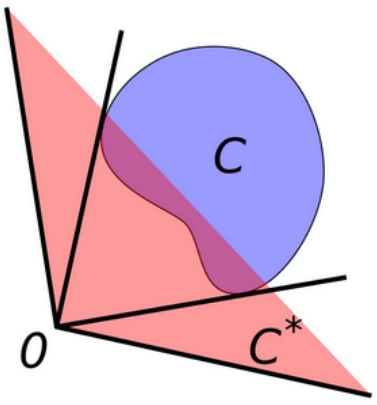


Figure 2.1: Illustration of the concept of the dual cone C^* of a set C

We now give several properties of dual cones that turn out to be useful later.

Proposition 2.2.1 (Statements taken from [7]) Let $P \subseteq \mathbb{R}^n$. The following holds for the dual cone P^*.

(i) P^* is a closed convex cone.

(ii) If P is closed, the interior of P^* is given by

$$\operatorname{int}(P^*) = \{y \in \mathbb{R}^n \mid \langle y, x\rangle > 0, \forall x \in P \setminus \{0\}\}.$$

(iii) If P is a closed convex cone, then $(P^*)^* = P$.

(iv) If $A \in \mathbb{R}^{n\times n}$ is an invertible matrix, then

$$(AP)^* = (A^\top)^{-1}P^*.$$

Proof. (i) We first show that P^* is closed. Let $x \in P$ and

$$H_x := \{y \in \mathbb{R}^n \mid \langle y, x\rangle \geq 0\}.$$

Clearly, for each fixed x, H_x is a closed half-space. Furthermore, by definition of P^*, we have

$$P^* = \bigcap_{x\in P} H_x.$$

Hence, P^* is the intersection of closed sets and so it is closed as well.

We now show that P^* is a convex cone. Let $x, y \in P^*$ and $\alpha, \beta \geq 0$. Fix $z \in P$. There holds

$$\begin{aligned}\langle \alpha x + \beta y, z\rangle &= \alpha\langle x, z\rangle + \beta\langle y, z\rangle \\ &\geq \alpha \cdot 0 + \beta \cdot 0 \\ &= 0.\end{aligned}$$

Since $z \in P$ was arbitrary, $\alpha x + \beta y \in P$ and by Proposition 2.1.1, P^* is a convex cone.

(ii) Let

$$\mathcal{P} := \{y \in \mathbb{R}^n \mid \langle y, x\rangle > 0, \forall x \in P \setminus \{0\}\}.$$

We will show that $\operatorname{int}(P^*) = \mathcal{P}$ by proving that $\operatorname{int}(P^*) \subseteq \mathcal{P}$ and $\mathcal{P} \subseteq \operatorname{int}(P^*)$.

Let $y \in \text{int}(P^*)$. This means that there exists a sufficiently small $\varepsilon > 0$ such that

$$B_\varepsilon(y) \subseteq P^*.$$

Assume by contradiction that $y \notin \mathcal{P}$. Since $y \in P^*$, there must exist $x \in P \setminus \{0\}$ such that $\langle y, x \rangle = 0$. Consider

$$z := y - \frac{\varepsilon}{2\|x\|} x.$$

Since

$$\|z - y\| = \frac{\varepsilon}{2\|x\|}\|x\| = \frac{\varepsilon}{2} < \varepsilon,$$

we have $z \in B_\varepsilon(y)$ but

$$\langle z, x \rangle = \langle y, x \rangle - \frac{\varepsilon}{2\|x\|}\langle x, x \rangle = -\frac{\varepsilon}{2}\|x\| < 0,$$

which is a contradiction since $z \in P^*$. Hence, $y \in \mathcal{P}$.

Let now $y \in \mathcal{P}$. Consider the set

$$X := P \cap S^{n-1}.$$

It is closed since it is the intersection of closed sets and bounded since it is a subset of the unit sphere. Hence, X is compact by the Heine–Borel theorem. Let

$$x^* := \arg\min_{x \in X} \langle y, x \rangle.$$

Note that x^* is well-defined by the extreme value theorem, since $\langle y, x \rangle$ is continuous in x and X is a compact set. Furthermore, $\langle y, x^* \rangle > 0$ because $y \in \mathcal{P}$ and $x^* \in P \setminus \{0\}$. Let now

$$z \in B_\varepsilon(y) \quad \text{where} \quad \varepsilon := \langle y, x^* \rangle.$$

For any $w \in P \setminus \{0\}$, there holds

$$\begin{aligned}
\langle z, w\rangle &= \langle y, w\rangle + \langle z - y, w\rangle \\
&= \|w\| \left(\left\langle y, \frac{w}{\|w\|} \right\rangle + \left\langle z - y, \frac{w}{\|w\|} \right\rangle \right) \\
&\geq \|w\| \left(\left\langle y, \frac{w}{\|w\|} \right\rangle - \|z - y\| \cdot \left\| \frac{w}{\|w\|} \right\| \right) \\
&= \|w\| \left(\left\langle y, \frac{w}{\|w\|} \right\rangle - \|z - y\| \right) \\
&> \|w\| \left(\left\langle y, \frac{w}{\|w\|} \right\rangle - \varepsilon \right) \\
&= \|w\| \left(\left\langle y, \frac{w}{\|w\|} \right\rangle - \langle y, x^*\rangle \right) \\
&\geq 0,
\end{aligned}$$

where the last inequality comes from the definition of x^*. Hence, $z \in P^*$ and since $z \in B_\varepsilon(y)$ was arbitrary, we conclude that $B_\varepsilon(y) \subseteq P^*$ and so $y \in \operatorname{int}(P^*)$.

(iii) See [7].

(iv) We have

$$\begin{aligned}
& z \in (AP)^* \\
& \Leftrightarrow z \in \{y \in \mathbb{R}^n \mid \langle y, x\rangle \geq 0, \forall x \in AP\} \\
& \Leftrightarrow z \in \{y \in \mathbb{R}^n \mid \langle y, Ax\rangle \geq 0, \forall x \in P\} \\
& \Leftrightarrow z \in \{y \in \mathbb{R}^n \mid \langle A^\top y, x\rangle \geq 0, \forall x \in P\} \\
& \Leftrightarrow A^\top z \in P^* \\
& \Leftrightarrow z \in (A^\top)^{-1} P^*,
\end{aligned}$$

as expected.

□

Example 2.2.1 (Dual cone of a linear subspace) Let $X \subseteq \mathbb{R}^n$ be a linear subspace. By definition of the dual cone, we have

$$X^* := \{y \in \mathbb{R}^n \mid \langle y, x\rangle \geq 0, \forall x \in X\}.$$

Clearly, $X^\perp \subseteq X^*$. Suppose there is a $y \in X^*$ such that $y \notin X^\perp$. This means that there exists $x \in X$ such that $\langle y, x\rangle > 0$. But since X is a linear subspace, $-x \in X$ and so we get

$$0 \leq \langle y, -x\rangle = -\langle y, x\rangle < 0,$$

which is impossible and so $X^* = X^\perp$. Hence, the dual cone of a linear subspace is its orthogonal complement.

Example 2.2.2 (Dual cone of a circular cone) From [35], for $\alpha \in [0, \pi/2]$, we have

$$\mathrm{Circ}_n(\alpha)^* = \mathrm{Circ}_n\left(\frac{\pi}{2} - \alpha\right).$$

Example 2.2.3 (Dual cone of an elliptical cone) As an elliptical cone is a linear transformation of a circular cone, by point (iv) of Proposition 2.2.1 and the previous example, the dual cone of an elliptical cone

$$\mathcal{A}\ \mathrm{Circ}_n(\alpha)$$

is given by

$$(\mathcal{A}^\top)^{-1}\ \mathrm{Circ}_n\left(\frac{\pi}{2} - \alpha\right),$$

where $\mathcal{A} \in \mathbb{R}^{n\times n}$ is a positive definite matrix and $\alpha \in [0, \pi/2]$.

2.3 Conic integral geometry

Conic integral geometry is a branch of mathematics that uses tools from convex geometry, probabilities and integration to better understand the geometric properties and transformations of convex sets, more precisely convex cones. While this is a vast subject, we are only interested in a small part of it, primarily quantifying the probability that a randomly rotated convex cone shares a ray with a fixed convex cone. For this, the following theorem is useful.

Theorem 2.3.1 (Conic kinematic formula) Let C and F be closed convex cones in $\mathbb{R}^n$ such that at most one of them is a linear subspace. Let Q be a $n \times n$ random orthogonal matrix drawn uniformly from the set of all $n \times n$ real orthogonal matrices. Then,

$$\mathbb{P}[QF \cap C \neq \{0\}] = \sum_{k=0}^{n} (1 + (-1)^{k+1}) \sum_{j=k}^{n} v_k(C) v_{n+k-j}(F),$$

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

Proof. See [28]. □

The conic kinematic formula therefore quantifies this probability of intersection in terms of quantities known as *conic intrinsic volumes* which are important geometric measures in the field of conic integral geometry.

2.3.1 Conic intrinsic volumes

Conic intrinsic volumes are geometric quantities that generalize the idea of area, volume, ... to convex sets that are not necessarily compact such as closed convex cones. For example, we are familiar with the concept of volume for a three-dimensional cube, but what about a circular cone? We could say that its volume is infinite, but that doesn't really provide any new information. This is where conic intrinsic volumes come into play. The formal definition of these quantities is quite complicated and is beyond the scope of this work. Hence, to better understand these concepts, we provide some of their main properties. For a closed convex cone $C \subseteq \mathbb{R}^n$,

- there are $n+1$ conic intrinsic volumes $v_0(C), v_1(C), \ldots, v_n(C)$,
- they form a probability distribution i.e.

$$\sum_{k=0}^{n} v_k(C) = 1 \quad \text{and} \quad v_k(C) \geq 0 \quad \text{for } k = 0, 1, \ldots, n$$

- and they are invariant under rotations i.e.

$$v_k(QC) = v_k(C)$$

  for any orthogonal matrix $Q \in \mathbb{R}^{n \times n}$.

This last property above is natural since geometric quantities such as area and volume should depend on the inherent structure of a shape rather than its orientation.

Unfortunately, explicit formulas for the conic intrinsic volumes are generally out of reach for arbitrary cones, except in the simplest cases such as linear subspaces or circular cones where symmetry allows for closed-form expressions.

Example 2.3.1 (From [10]) The conic intrinsic volumes of a d-dimensional linear subspace $\mathcal{L}_d \subseteq \mathbb{R}^n$ are given by

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

Example 2.3.2 (From [10]) The conic intrinsic volumes of a circular cone with angle $\alpha \in [0, \frac{\pi}{2}]$ are given by

$$v_k(\mathrm{Circ}_n(\alpha)) = \frac{1}{2}\binom{(n-2)/2}{(k-1)/2} \sin^{k-1}(\alpha)\cos^{n-k-1}(\alpha) \quad \text{for } k = 1, 2, \ldots, n-1,$$

where

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

Furthermore,

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

and

$$v_n(\mathrm{Circ}_n(\alpha)) = \frac{n-1}{2}\binom{(n-2)/2}{(n-1)/2}\int_0^{\alpha} \sin^{n-2}(x)\,dx.$$

The conic kinematic formula coupled with Example 2.3.1 enables us to quantify the probability that a randomly rotated linear subspace (which is indeed a closed convex cone) shares a ray with a fixed convex cone.

Corollary 2.3.1 (Crofton formula) Let C be a closed convex cone in $\mathbb{R}^n$ and $\mathcal{L}_d$ be a d-dimensional linear subspace. Let Q be a $n \times n$ random orthogonal matrix drawn uniformly from the set of all $n \times n$ real orthogonal matrices. We have

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

with

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

Proof. Just apply Theorem 2.3.1 with $F = \mathcal{L}_d$, using Example 2.3.1. □

2.3.2 Statistical dimension

Although we don't have workable expressions for the conic intrinsic volumes of arbitrary cones, a general related result has been shown in [6] for closed convex cones. That is, for every closed convex, the distribution of conic intrinsic volumes concentrates sharply around its mean value. As an example, the distribution of these volumes for the circular cone with angle $\pi/6$ is shown in the figure below.

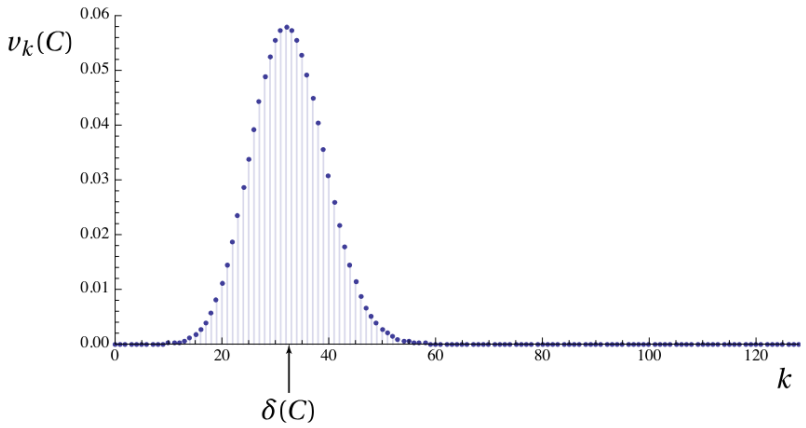


Figure 2.2: (From [6]) Concentration of the conic intrinsic volumes of $\mathrm{Circ}_{128}(\pi/6)$

We see that the distribution concentrates around the mean value $\delta(C) \approx 32.5$. From these results, it is therefore natural to make the following definition.

Definition 2.3.1 Let $C \subseteq \mathbb{R}^n$ be a closed convex cone. The statistical dimension $\delta(C)$ of the cone is defined as

$$\delta(C) := \sum_{k=0}^{n} k v_k(C).$$

The name "statistical dimension" was not chosen arbitrarily, as illustrated by the example below.

Example 2.3.3 Let $\mathcal{L}_d \subseteq \mathbb{R}^n$ be a d-dimensional linear subspace. Then,

$$\delta(\mathcal{L}_d) = \sum_{k=0}^{n} k v_k(\mathcal{L}_d) = d$$

where we used Example 2.3.1 for the conic intrinsic volumes of a linear subspace.

Hence, for linear subspaces, the statistical dimension is just the dimension of the subspace. More generally, the statistical dimension can be viewed as a natural extension of the dimension of a linear subspace to convex cones.

Example 2.3.4 (From [6]) The statistical dimension of a circular cone with angle $\alpha \in [0, \pi/2]$ is given by

$$\delta(\mathrm{Circ}_n(\alpha)) = n \cdot \frac{\Gamma(n/2)}{\sqrt{\pi}\Gamma((n-1)/2)} \int_0^{\pi} \sin^{n-2}(\beta) F(\beta)\, d\beta,$$

where

$$F(\beta) := \begin{cases} 1, & 0 \le \beta < \alpha, \\ \cos^2(\beta - \alpha), & \alpha \le \beta < \frac{\pi}{2} + \alpha, \\ 0, & \frac{\pi}{2} + \alpha \le \beta \le \pi. \end{cases}$$

Definition 2.3.1 is known as the intrinsic formulation of the statistical dimension, meaning that it is expressed in terms of the conic intrinsic volumes. It turns out that there exists another equivalent formulation which is often easier to deal with, especially in a numerical point of view.

Theorem 2.3.2 (Gaussian formulation of the statistical dimension) For a closed convex cone $C \subseteq \mathbb{R}^n$, there holds

$$\delta(C) = \mathbb{E}\left[\|\Pi_C(g)\|^2\right],$$

where $g \sim \mathcal{N}(0, I_n)$ and

$$\Pi_C(x) := \operatorname*{arg\,min}_{y \in C} \|x - y\|^2$$

is the Euclidean projection of a vector $x \in \mathbb{R}^n$ onto the cone C.

Proof. See [6]. □

With the help of Theorem 2.3.2, we now have a way to numerically compute the statistical dimension of arbitrary closed convex cones. In fact, it suffices to generate standard normal random vectors and compute the average squared norm of their projections onto C, which amounts to solving a convex optimization problem.

As mentioned earlier, the statistical dimension is generally more tractable than the more complicated conic intrinsic volumes. It is therefore natural to seek an analog of the conic kinematic formula directly expressed in terms of the statistical dimension. By leveraging the concentration of conic intrinsic volumes around the statistical dimension, the conic kinematic formula can indeed be simplified, as illustrated in the theorem below. For this, define

$$\delta_{\min}(C) := \min(\delta(C), \delta(C^*))$$

for a closed convex cone $C \subseteq \mathbb{R}^n$ and the function

$$p_C(\lambda) := 4 \exp\left(\frac{-\lambda^2/8}{\delta_{\min}(C) + \lambda}\right)$$

for $\lambda \geq 0$.

Theorem 2.3.3 (Approximate kinematic formula) Let C and K be closed convex cones in $\mathbb{R}^n$ and draw a random orthogonal matrix $Q \in \mathbb{R}^{n \times n}$. Then, for any $\lambda \geq 0$, it holds that

$$\delta(C) + \delta(K) \leq n - 2\lambda \implies \mathbb{P}[C \cap QK \neq \{0\}] \leq p_C(\lambda) + p_K(\lambda);$$

$$\delta(C) + \delta(K) \geq n + 2\lambda \implies \mathbb{P}[C \cap QK \neq \{0\}] \geq 1 - (p_C(\lambda) + p_K(\lambda)).$$

Proof. See [6]. □

When the closed convex cone K is a subspace, we have the following simplified version of the Crofton formula.

Theorem 2.3.4 (Approximate Crofton formula) Let C be a closed convex cone in $\mathbb{R}^n$ and $\mathcal{L}_d$ be a d-dimensional linear subspace. Let Q be a $n \times n$ random orthogonal matrix drawn uniformly from the set of all $n \times n$ real orthogonal matrices. For any $\lambda \geq 0$, there holds

$$n - d \geq \delta(C) + \lambda \implies \mathbb{P}[C \cap Q\mathcal{L}_d \neq \{0\}] \leq p_C(\lambda);$$

$$n - d \leq \delta(C) - \lambda \implies \mathbb{P}[C \cap Q\mathcal{L}_d \neq \{0\}] \geq 1 - p_C(\lambda).$$

Proof. See [6]. □

2.4 Gaussian width

In high-dimensional probability, an important quantity called *Gaussian width* is often encountered.

Definition 2.4.1 The Gaussian width of a subset $T \subseteq \mathbb{R}^n$ is defined as

$$w(T) := \mathbb{E}[\sup_{x \in T} \langle g, x \rangle] \quad \text{where } g \sim \mathcal{N}(0, I_n).$$

As for the conic intrinsic volumes, the Gaussian width is a basic geometric quantity associated with a set T such as area or volume. Before diving into a more precise geometric interpretation, we consider next the main properties of the Gaussian width.

Proposition 2.4.1 (Statements taken from [31]) Let $T, S \subseteq \mathbb{R}^n$ be non-empty. The following holds for the Gaussian width.

(i) The Gaussian width is invariant under affine unitary transformations i.e.

$$w(UT + y) = w(T)$$

for every orthogonal matrix $U \in \mathbb{R}^{n \times n}$ and any vector $y \in \mathbb{R}^n$.

(ii) We have

$$w(T + S) = w(T) + w(S) \quad \text{and} \quad w(aT) = |a| w(T)$$

for any $a \in \mathbb{R}$.

(iii) We have

$$w(T) = \frac{1}{2}w(T-T) = \frac{1}{2}\mathbb{E}[\sup_{x,y\in T}\langle g, x-y\rangle].$$

(iv) We have

$$\frac{1}{\sqrt{2\pi}}\cdot \operatorname{diam}(T) \le w(T) \le \frac{\sqrt{n}}{2}\cdot \operatorname{diam}(T)$$

where

$$\operatorname{diam}(T) := \sup_{x,y\in T}\|x-y\|$$

is the diameter of T.

Proof. (i) Since the normal distribution is invariant under rotations, there holds

$$\begin{aligned} w(UT+y) &= \mathbb{E}[\sup_{z\in UT+y}\langle g, z\rangle] \\ &= \mathbb{E}[\sup_{x\in T}\langle g, Ux+y\rangle] \\ &= \mathbb{E}[\sup_{x\in T}\langle g, Ux\rangle] + \mathbb{E}[\sup_{x\in T}\langle g, y\rangle] \\ &= \mathbb{E}[\sup_{x\in T}\langle U^\top g, x\rangle] + \mathbb{E}[\langle g, y\rangle] \\ &= \mathbb{E}[\sup_{x\in T}\langle U^\top g, x\rangle] \\ &= w(T). \end{aligned}$$

(ii) We have

$$\begin{aligned} w(T+S) &= \mathbb{E}[\sup_{z\in T+S}\langle g, z\rangle] \\ &= \mathbb{E}[\sup_{x\in T, y\in S}\langle g, x+y\rangle] \\ &= \mathbb{E}[\sup_{x\in T}\langle g, x\rangle] + \mathbb{E}[\sup_{y\in S}\langle g, y\rangle] \\ &= w(T) + w(S) \end{aligned}$$

,

$$\begin{aligned} w(aT) &= \mathbb{E}[\sup_{z\in aT}\langle g, z\rangle] \\ &= \mathbb{E}[\sup_{x\in T}\langle g, ax\rangle] \\ &= a\mathbb{E}[\sup_{x\in T}\langle g, x\rangle] \\ &= aw(T) \end{aligned}$$

if $a \geq 0$ and by symmetry of the normal distribution,

$$\begin{aligned} w(aT) &= \mathbb{E}[\sup_{z \in aT} \langle g, z \rangle] \\ &= \mathbb{E}[\sup_{x \in T} \langle g, ax \rangle] \\ &= \mathbb{E}[\sup_{x \in T} \langle -g, -ax \rangle] \\ &= -a\mathbb{E}[\sup_{x \in T} \langle -g, x \rangle] \\ &= -a\mathbb{E}[\sup_{x \in T} \langle g, x \rangle] \\ &= -aw(T) \end{aligned}$$

otherwise. Hence, $w(aT) = |a|w(T)$ as expected.

(iii) By (iii), there holds

$$w(T) = \frac{1}{2}(w(T) + w(T)) = \frac{1}{2}(w(T) + w(-T)) = \frac{1}{2}w(T - T)$$

(iv) See [31].

□

Property (iii) turns out to be very useful in understanding what the Gaussian width represents geometrically. In fact, the Gaussian width of a set $T \subseteq \mathbb{R}^n$ is actually a scaled version of the expected value of the smallest distance between two parallel hyperplanes orthogonal to a random unit vector $\theta \subseteq S^{n-1}$ and that enclose the set T. This is illustrated in Figure 2.3.

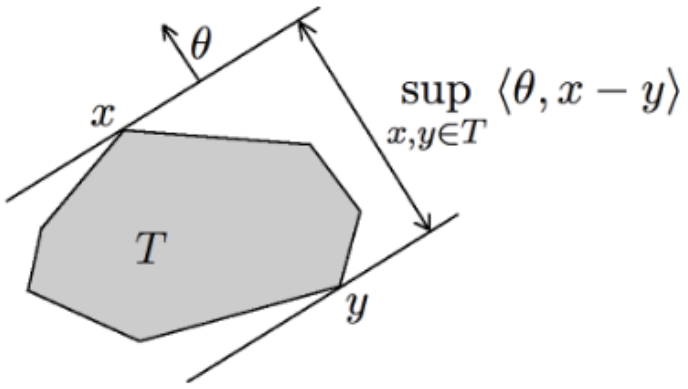


Figure 2.3: The width of a set $T \subseteq \mathbb{R}^n$ in the direction of a unit vector θ

It is easy to see that the width of T in the direction of θ can be expressed analytically as

$$\sup_{x,y \in T} \langle \theta, x - y \rangle$$

and so the expected value of this width is given by

$$w_s(T - T)$$

where

$$w_s(T) := \mathbb{E}[\sup_{x \in T} \langle \theta, x \rangle] \quad \text{with } \theta \sim \text{Uniform}(S^{n-1}).$$

The quantity $w_s(T)$ is known in the literature as the spherical width of T. The difference from the Gaussian width is in the way we take the expectation. For the spherical width, the random vectors are uniform whereas for the Gaussian width, they are normal. It turns out that the latter is a scaled version of the former, which we prove in the following lemma.

Lemma 2.4.1 (Inspired from [31]) Let $T \subseteq \mathbb{R}^n$. The Gaussian width and the spherical width are related in the following way:

$$w(T) = \sqrt{2}\frac{\Gamma((n+1)/2)}{\Gamma(n/2)} w_s(T)$$

Before proving Lemma 2.4.1, we will need the proposition below.

Proposition 2.4.2 (Statement taken from [31]) Let $g \sim \mathcal{N}(0, I_n)$ and write g in polar form i.e.

$$g = r\theta$$

where $r = \|g\|$ and $\theta = g/\|g\|$. Then r and θ are independent random variables, $r \sim \chi_n$ and $\theta \sim \text{Uniform}(S^{n-1})$.

Proof. The probability density function of g is given by

$$f_g(x) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\|x\|^2\right).$$

Making the change of variables

$$x = \rho\phi$$

with $\rho \in \mathbb{R}^+$ and $\phi \in S^{n-1}$, the joint density of (r, θ) becomes

$$f_{r,\theta}(\rho, \phi) = f_g(\rho\phi)\rho^{n-1}$$

since $dx = \rho^{n-1} d\rho \, d\phi$ and so

$$\begin{aligned} f_{r,\theta}(\rho, \phi) &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{\rho^2}{2}\right) \rho^{n-1} \\ &= \left(\frac{1}{2^{(n/2)-1}\Gamma(n/2)} \exp\left(-\frac{\rho^2}{2}\right) \rho^{n-1}\right) \cdot \left(\frac{\Gamma(n/2)}{2\pi^{n/2}}\right) \\ &= f_r(\rho) f_\theta(\phi) \end{aligned}$$

where $f_r(\cdot)$ is the PDF of a chi distribution and $f_\theta(\cdot)$ the PDF of a uniform distribution. Hence, r and θ are independent with $r \sim \chi_n$ and $\theta \sim \text{Uniform}(S^{n-1})$, as expected. □

Proof. (Lemma 2.4.1) Let $g \sim \mathcal{N}(0, I_n)$ be written in polar form i.e.

$$g = r\theta$$

with $r = \|g\|$ and $\theta = g/\|g\|$. With the help of Proposition 2.4.2, there holds

$$\begin{aligned} w(T) &= \mathbb{E}[\sup_{x \in T} \langle r\theta, x \rangle] \\ &= \mathbb{E}[r \sup_{x \in T} \langle \theta, x \rangle] \\ &= \mathbb{E}[r] \cdot \mathbb{E}[\sup_{x \in T} \langle \theta, x \rangle] \\ &= \sqrt{2} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} w_s(T) \end{aligned}$$

where the coefficient of $w_s(T)$ comes from the known mean value of the chi distribution with n degrees of freedom. □

We now compute the Gaussian width for some of the most common geometric shapes.

Example 2.4.1 The Gaussian width of the unit sphere is

$$w(S^{n-1}) = \mathbb{E}[\sup_{x \in S^{n-1}} \langle g, x \rangle].$$

For each fixed $g \in \mathbb{R}^n$, we have by the Cauchy-Schwarz inequality

$$\sup_{x \in S^{n-1}} \langle g, x \rangle \leq \|g\| \sup_{x \in S^{n-1}} \|x\| = \|g\|.$$

Picking $x = g/\|g\| \in S^{n-1}$ gives

$$\langle g, g/\|g\| \rangle = \|g\|$$

Hence,

$$\sup_{x \in S^{n-1}} \langle g, x \rangle = \|g\|$$

and so

$$w(S^{n-1}) = \mathbb{E}[\|g\|] = \sqrt{2} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)}.$$

Example 2.4.2 The Gaussian width of the Euclidean unit ball is

$$w(B^n) = \mathbb{E}[\sup_{x \in B^n} \langle g, x \rangle].$$

For each fixed $g \in \mathbb{R}^n$, similarly as in the previous example, we have

$$\sup_{x \in B^n} \langle g, x \rangle \leq \|g\| \sup_{x \in B^n} \|x\| = \|g\|$$

and picking $x = g/\|g\| \in B^n$ makes the inequality tight. Hence,

$$w(B^n) = w(S^{n-1}) = \sqrt{2}\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}.$$

In the previous section, we introduced the statistical dimension as a more tractable alternative to the conic intrinsic volumes. We were even able to quantify the probability that a randomly rotated linear subspace shares a ray with a fixed cone, using only this quantity. The Gaussian width also appears to offer similar results, but it relates instead the Gaussian width of a set to the probability that a randomly rotated linear subspace avoids intersecting a fixed cone, through Gordon's theorem.

Theorem 2.4.1 (Gordon's escape through a mesh theorem) Let $T \subseteq \mathbb{R}^n$ be a closed subset of the unit sphere S^{n-1}. If

$$w(T) < \sqrt{d},$$

then

$$\mathbb{P}[Y \cap T = \emptyset] \geq 1 - \frac{5}{2}\exp\left(-\frac{(d/\sqrt{d+1} - w(T))^2}{18}\right)$$

where Y is a random $(n-d)$-dimensional linear subspace.

Proof. See [30] and [14]. □

As we saw previously, several geometric shapes were expressed as linear transformations of simpler ones (for example, an elliptical cone can be viewed as a linear transformation of a circular cone). Therefore, it makes sense to study how the Gaussian width behaves under linear transformations. This is formalized in the following theorem.

Theorem 2.4.2 Let $C \subseteq \mathbb{R}^n$ be a closed convex cone. Then, for an invertible matrix $A \in \mathbb{R}^{n \times n}$, there holds

$$\frac{1}{\kappa(A)} w(C \cap S^{n-1}) \leq w(AC \cap S^{n-1}) \leq \kappa(A) w(C \cap S^{n-1})$$

where $\kappa(A) = \|A^{-1}\| \cdot \|A\|$ is the condition number of A.

To prove the previous theorem, we need the following definitions and lemma.

Definition 2.4.2 (Random process) A random process is a collection of random variables $(X_t)_{t\in T}$ on the same probability space, which are indexed by elements t of some set $T \subseteq \mathbb{R}^n$.

Definition 2.4.3 (Gaussian process) Let $T \subseteq \mathbb{R}^n$. A random process $(X_t)_{t\in T}$ is called a Gaussian process if, for any finite subset $T_0 \subseteq T$, the random vector $(X_t)_{t\in T_0}$ has normal distribution.

Lemma 2.4.2 (Sudakov-Fernique's inequality) Let $T \subseteq \mathbb{R}^n$. Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two mean zero Gaussian processes. Assume that for all $t, s \in T$, we have

$$\mathbb{E}[(X_t - X_s)^2] \leq \mathbb{E}[(Y_t - Y_s)^2].$$

Then,

$$\mathbb{E}[\sup_{t\in T} X_t] \leq \mathbb{E}[\sup_{t\in T} Y_t].$$

Proof. See [31]. □

Proof. (Theorem 2.4.2) We first show the right inequality. By definition of the Gaussian width, we have

$$w(AC \cap S^{n-1}) = \mathbb{E}\left[\sup_{y\in AC\cap S^{n-1}} \langle y, g\rangle\right].$$

Let

$$X_t := \langle t, g\rangle \quad \text{and} \quad Y_t := \|A^{-1}\| \cdot \left\langle \frac{t}{\|A^{-1}t\|}, g \right\rangle$$

for $t \in AC \cap S^{n-1}$. Note that

$$\langle t, g\rangle = \sum_{i=1}^{n} t_i g_i \sim \mathcal{N}(0, \|t\|^2)$$

since $\langle t, g\rangle$ is a weighted sum of standard normal variables. Hence, (X_t) and (Y_t) are Gaussian processes with

$$\mathbb{E}[X_t] = \mathbb{E}[Y_t] = 0.$$

Furthermore, for all $t, s \in AC \cap S^{n-1}$, there holds

$$\mathbb{E}[(X_t - X_s)^2] = \mathbb{E}[\langle t - s, g\rangle^2] = \mathbb{V}[\langle t - s, g\rangle] + \mathbb{E}[\langle t - s, g\rangle]^2 = \|t - s\|^2$$

and similarly,

$$\mathbb{E}[(Y_t - Y_s)^2] = \|A^{-1}\|^2 \cdot \mathbb{E}\left[\left\langle \frac{t}{\|A^{-1}t\|} - \frac{s}{\|A^{-1}s\|}, g \right\rangle^2\right] = \|A^{-1}\|^2 \cdot \left\| \frac{t}{\|A^{-1}t\|} - \frac{s}{\|A^{-1}s\|} \right\|^2.$$

In order to apply the Sudakov-Fernique's inequality, we need to show that

$$\|t - s\| \leq \|A^{-1}\| \cdot \left\| \frac{t}{\|A^{-1}t\|} - \frac{s}{\|A^{-1}s\|} \right\|.$$

Note that

$$\|A^{-1}t\| \leq \|A^{-1}\| \cdot \|t\| = \|A^{-1}\|$$

since $t \in S^{n-1}$ and so

$$\frac{\|t-s\|}{\|A^{-1}\|} = \frac{\|t-s\|}{2}\left(\frac{1}{\|A^{-1}\|} + \frac{1}{\|A^{-1}\|}\right) \leq \frac{\|t-s\|}{2}\left(\frac{1}{\|A^{-1}t\|} + \frac{1}{\|A^{-1}s\|}\right).$$

Hence, it suffices to show that

$$\|t-s\|\left(\frac{1}{\|A^{-1}t\|} + \frac{1}{\|A^{-1}s\|}\right) \leq 2\left\| \frac{t}{\|A^{-1}t\|} + \frac{s}{\|A^{-1}s\|} \right\|.$$

For simplicity, let

$$a := \|A^{-1}t\| \quad \text{and} \quad b := \|A^{-1}s\|.$$

We have

$$\begin{aligned}
&\|t-s\|\left(\frac{1}{a} + \frac{1}{b}\right) \leq 2\left\|\frac{t}{a} + \frac{s}{b}\right\| \\
&\Leftrightarrow \|t-s\|\left(\frac{a+b}{ab}\right) \leq 2\left\|\frac{bt-as}{ab}\right\| \\
&\Leftrightarrow \|t-s\|^2(a+b)^2 \leq 4\|bt-as\|^2 \\
&\Leftrightarrow (\|t\|^2 + \|s\|^2 - 2\langle t,s\rangle)(a+b)^2 \leq 4(b^2\|t\|^2 + a^2\|s\|^2 - 2ab\langle t,s\rangle) \\
&\Leftrightarrow (2 - 2\langle t,s\rangle)(a+b)^2 \leq 4(a^2+b^2-2ab\langle t,s\rangle) \\
&\Leftrightarrow (1 - \langle t,s\rangle)(a+b)^2 \leq 2(a^2+b^2-2ab\langle t,s\rangle) \\
&\Leftrightarrow 2(a^2+b^2) - (a+b)^2 + ((a+b)^2 - 4ab)\langle t,s\rangle \geq 0 \\
&\Leftrightarrow a^2+b^2-2ab + (a^2+b^2-2ab)\langle t,s\rangle \geq 0 \\
&\Leftrightarrow (a-b)^2 + (a-b)^2\langle t,s\rangle \geq 0 \\
&\Leftrightarrow (a-b)^2(1+\langle t,s\rangle) \geq 0.
\end{aligned}$$

Note that, by the Cauchy-Schwarz inequality, there holds

$$-\langle t,s\rangle = \langle -t,s\rangle \leq \|-t\| \cdot \|s\| = 1$$

and so we indeed have

$$(a-b)^2(1+\langle t,s\rangle)\geq 0.$$

Therefore, this shows that

$$\mathbb{E}[(X_t-X_s)^2]\leq \mathbb{E}[(Y_t-Y_s)^2]$$

for all $t,s\in AC\cap S^{n-1}$ and so we can apply the Sudakov-Fernique's inequality. Hence,

$$\begin{aligned}
w(AC\cap S^{n-1}) &= \mathbb{E}\left[\sup_{y\in AC\cap S^{n-1}}\langle y,g\rangle\right]\\
&= \mathbb{E}\left[\sup_{t\in AC\cap S^{n-1}} X_t\right]\\
&\leq \mathbb{E}\left[\sup_{t\in AC\cap S^{n-1}} Y_t\right]\\
&= \|A^{-1}\|\cdot\mathbb{E}\left[\sup_{y\in AC\cap S^{n-1}}\left\langle\frac{y}{\|A^{-1}y\|},g\right\rangle\right]\\
&= \|A^{-1}\|\cdot\mathbb{E}\left[\sup_{x\in C\cap A^{-1}S^{n-1}}\left\langle A\frac{x}{\|x\|},g\right\rangle\right]\\
&\leq \|A^{-1}\|\cdot\mathbb{E}\left[\sup_{x\in C\cap S^{n-1}}\langle Ax,g\rangle\right],
\end{aligned}$$

where the last inequality comes from the fact that, if

$$x\in C\cap A^{-1}S^{n-1},$$

then

$$\frac{x}{\|x\|}\in C\cap S^{n-1}$$

because $x/\|x\|\in S^{n-1}$ and $x/\|x\|\in C$ since C is a cone. We now let

$$X_t:=\langle At,g\rangle\quad\text{and}\quad Y_t:=\|A\|\cdot\langle t,g\rangle$$

for $t\in C\cap S^{n-1}$. Again, (X_t) and (Y_t) are Gaussian processes with

$$\mathbb{E}[X_t]=\mathbb{E}[Y_t]=0.$$

Furthermore,

$$\mathbb{E}[(X_t-X_s)^2]=\mathbb{E}[\langle A(t-s),g\rangle^2]=\|A(t-s)\|^2$$

and

$$\mathbb{E}[(Y_t-Y_s)^2]=\|A\|^2\cdot\mathbb{E}[\langle t-s,g\rangle^2]=\|A\|^2\cdot\|t-s\|^2.$$

Clearly,

$$\mathbb{E}[(X_t - X_s)^2] \leq \mathbb{E}[(Y_t - Y_s)^2]$$

and so, by the Sudakov-Fernique's inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{x \in C \cap S^{n-1}} \langle Ax, g \rangle\right] &= \mathbb{E}\left[\sup_{t \in C \cap S^{n-1}} X_t\right] \\
&\leq \mathbb{E}\left[\sup_{t \in C \cap S^{n-1}} Y_t\right] \\
&= \|A\| \cdot \mathbb{E}\left[\sup_{x \in C \cap S^{n-1}} \langle x, g \rangle\right] \\
&= \|A\| \cdot w(C \cap S^{n-1}).
\end{aligned}$$

Hence,

$$w(AC \cap S^{n-1}) \leq \|A^{-1}\| \cdot \|A\| \cdot w(C \cap S^{n-1}) = \kappa(A) \cdot w(C \cap S^{n-1})$$

and this shows the right inequality of the theorem.

To prove the left inequality, note that

$$w(C \cap S^{n-1}) = w(A^{-1}AC \cap S^{n-1}) \leq \kappa(A^{-1}) w(AC \cap S^{n-1}).$$

Since $\kappa(A^{-1}) = \kappa(A)$, this ends the proof. □

Note that, as for the statistical dimension, the Gaussian width can be expressed as the expectation of the norm of the Euclidean projection of Gaussian random variables, which can be useful for a numerical computation of the Gaussian width. This is made explicit in the proposition below.

Proposition 2.4.3 Let $C \subseteq \mathbb{R}^n$ be a closed convex cone. Then,

$$w(C \cap S^{n-1}) = \mathbb{E}\left[\|\Pi_C(g)\|\right],$$

where

$$\Pi_C(x) := \arg\min_{y \in C} \|x - y\|^2$$

is the Euclidean projection of a vector $x \in \mathbb{R}^n$ onto the cone C.

Proof. See [34]. □

To conclude this section, we emphasize that the Gaussian width and the statistical dimension are closely related, as stated in the following proposition.

Proposition 2.4.4 Let $C \subseteq \mathbb{R}^n$ be a closed convex cone. Then,

$$w^2(C \cap S^{n-1}) \leq \delta(C) \leq w^2(C \cap S^{n-1}) + 1.$$

Proof. See [6]. □

Chapter 3

Functions with anisotropic dimensionality

3.1 Definition and motivation

Previously, we saw an expression (the bound (1.1)) for the lower bound on the probability that a randomized reduced problem, constructed using random embeddings, yields an ε-minimizer of the original global optimization problem, under the assumption that the objective function is Lipschitz continuous. To tighten this bound, we focus on a more structured class of functions, which we refer to as functions with anisotropic dimensionality, or AD functions.

Definition 3.1.1 (AD function) A function $f : \mathbb{R}^D \to \mathbb{R}$ is said to have anisotropic dimensionality d_a, with $d_a \leq D$, if

- there exists a linear subspace $\mathcal{T} \subseteq \mathbb{R}^D$ of dimension d_a such that

  $$|f(x+\eta) - f(x)| \leq L_1\|\eta\| \quad \text{for all } \eta \in \mathcal{T}$$

  and

  $$|f(x+\xi) - f(x)| \leq L_2\|\xi\| \quad \text{for all } \xi \in \mathcal{T}^\perp$$

  with $L_1 \gg L_2$ and $L_2 \ll 1$;

- d_a is the smallest integer with this property.

Intuitively, AD functions vary primarily along a linear subspace $\mathcal{T}$, with small variation along the orthogonal subspace $\mathcal{T}^\perp$. See Figure 3.1 for an illustration.

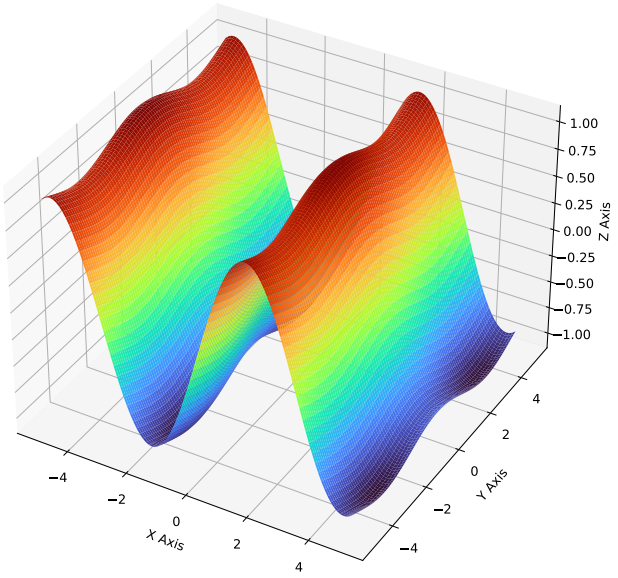


Figure 3.1: Function with anisotropic dimensionality $d_a = 1$. Here, $f(x, y) = \sin(x) + 0.1\cos(y)$ with $L_1 = 1$ and $L_2 = 0.1$.

Note that AD functions are Lipschitz continuous functions, as stated in the following proposition.

Proposition 3.1.1 Let $f : \mathbb{R}^D \to \mathbb{R}$ be an AD function. Then f is Lipschitz continuous with Lipschitz constant $L_1 + L_2$. That is,

$$|f(x) - f(y)| \leq (L_1 + L_2)\|x - y\|$$

for all $x, y \in \mathbb{R}^D$.

Proof. Let $x, y \in \mathbb{R}^D$ be arbitrary. Since $\mathbb{R}^D = \mathcal{T} \oplus \mathcal{T}^\perp$, we can decompose y into $y = x + \eta + \xi$ with $\eta \in \mathcal{T}$ and $\xi \in \mathcal{T}^\perp$. Then, by the triangular inequality and the definition of an AD function, there holds

$$\begin{aligned}
|f(x) - f(y)| &= |f(x) - f(x + \eta + \xi)| \\
&= |f(x) - f(x + \eta) + f(x + \eta) - f(x + \eta + \xi)| \\
&\leq |f(x + \eta) - f(x)| + |f(x + \eta + \xi) - f(x + \eta)| \\
&\leq L_1\|\eta\| + L_2\|\xi\|.
\end{aligned}$$

Next, we show that

$$L_1\|\eta\| + L_2\|\xi\| \leq (L_1 + L_2)\|\eta + \xi\|.$$

Indeed, using the Cauchy-Schwarz inequality and the orthogonality of η and ξ, we have

$$\begin{aligned} L_1\|\eta\| + L_2\|\xi\| &\leq \sqrt{L_1^2 + L_2^2}\sqrt{\|\eta\|^2 + \|\xi\|^2} \\ &= \sqrt{L_1^2 + L_2^2}\|\eta + \xi\| \\ &= \|(L_1, L_2)^\top\| \cdot \|\eta + \xi\| \\ &= \|(L_1, 0)^\top + (0, L_2)^\top\| \cdot \|\eta + \xi\| \\ &\leq (\|(L_1, 0)^\top\| + \|(0, L_2)^\top\|)\|\eta + \xi\| \\ &= (L_1 + L_2)\|\eta + \xi\| \end{aligned}$$

where we used the triangular inequality in the last inequality. Hence,

$$|f(x) - f(y)| \leq (L_1 + L_2)\|\eta + \xi\| = (L_1 + L_2)\|x - y\|,$$

which shows that f is Lipschitz continuous with Lipschitz constant $L_1 + L_2$. □

3.2 Lower bounds

3.2.1 Problem definition and assumptions

We consider the global optimization problem

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

where $f : \mathcal{X} \to \mathbb{R}$ is a function with anisotropic dimensionality d_a and $\mathcal{X} \subseteq \mathbb{R}^D$ is a set with non-empty interior. The randomized reduced problem is then given by

$$\begin{aligned} \min_{y \in \mathbb{R}^d} \quad & f(Ay + p) \\ \text{s.t.} \quad & Ay + p \in \mathcal{X}, \end{aligned} \tag{RP$\mathcal{X}$}$$

where $A \in \mathbb{R}^{D \times d}$ is a Gaussian random matrix and $p \in \mathcal{X}$ is arbitrary or user-defined.

We are interested in the probability that (RP$\mathcal{X}$) is ε-successful.

Definition 3.2.1 (From [10]) (RP$\mathcal{X}$) is ε-successful if there exists $y \in \mathbb{R}^d$ such that $Ay + p \in \mathcal{X}$ and $f(Ay + p) \leq f^* + \varepsilon$, where $\varepsilon > 0$ is a desired accuracy tolerance.

Given a desired accuracy $\varepsilon > 0$, we define the set of ε-minimizers of (P) as

$$G_\varepsilon := \{x \in \mathcal{X} : f(x) \leq f^* + \varepsilon\}.$$

Throughout the remainder of this thesis, we will operate under the following assumptions, which are necessary to ensure the validity of our results.

Assumption 3.2.1 (Inspired from [10]) The function $f : \mathcal{X} \to \mathbb{R}$ has anisotropic dimensionality d_a, with $\mathcal{T}$ and $\mathcal{T}^\perp$ spanned by the columns of orthonormal matrices $U \in \mathbb{R}^{D\times d_a}$ and $V \in \mathbb{R}^{D\times(D-d_a)}$ respectively. The unique Euclidean projections of any vector $x \in \mathcal{X}$ onto $\mathcal{T}$ and $\mathcal{T}^\perp$ are then respectively given by $UU^\top x$ and $VV^\top x$.

Assumption 3.2.2 There exists a global minimizer x^* of (P) such that $\mathcal{E}_E(x^*) \subset \mathcal{X}$, where the rotated ellipsoid $\mathcal{E}_E(x^*)$ is defined as

$$\mathcal{E}_E(x^*) := \{x \in \mathbb{R}^D : (x - x^\star)^\top E(x - x^\star) \leq 1\},$$

where $E = Q\Lambda Q^\top$ with

$$Q = \begin{bmatrix} U & V \end{bmatrix} \quad \text{and} \quad \Lambda = \begin{pmatrix} \Lambda_1 & \\ & \Lambda_2 \end{pmatrix},$$

where

$$\Lambda_1 = \operatorname{diag}\left(1/l_1^2\right) \in \mathbb{R}^{d_a\times d_a} \quad \text{and} \quad \Lambda_2 = \operatorname{diag}\left(1/l_2^2\right) \in \mathbb{R}^{(D-d_a)\times(D-d_a)}$$

with

$$l_1 = \frac{\varepsilon}{2L_1} \quad \text{and} \quad l_2 = \frac{\varepsilon}{2L_2}$$

being half the length of the principal axes of the ellipsoid.

Remark The matrix $Q \in \mathbb{R}^{D\times D}$ is orthonormal since U and V are orthonormal matrices.

The choice of the length of the semi-axis of the ellipsoid in Assumption 3.2.2 is not arbitrary, as illustrated in the following proposition.

Proposition 3.2.1 Let $\mathcal{E}_E(x^*)$ be the ellipsoid defined in Assumption 3.2.2. Then,

$$\mathcal{E}_E(x^*) \subseteq G_\varepsilon.$$

Proof. Let $x \in \mathcal{E}_E(x^*)$ and decompose x into $x = x^\star + \eta + \xi$ for $\eta \in \mathcal{T}$ and $\xi \in \mathcal{T}^\perp$.

There holds

$$\begin{aligned}
(x - x^\star)^\top E(x - x^\star) &= (\eta + \xi)^\top Q \Lambda Q^\top (\eta + \xi) \\
&= (\eta + \xi)^\top \begin{bmatrix} U & V \end{bmatrix} \begin{pmatrix} \Lambda_1 & \\ & \Lambda_2 \end{pmatrix} \begin{bmatrix} U^\top (\eta + \xi) \\ V^\top (\eta + \xi) \end{bmatrix} \\
&= (\eta + \xi)^\top \begin{bmatrix} U & V \end{bmatrix} \begin{bmatrix} \Lambda_1 U^\top (\eta + \xi) \\ \Lambda_2 V^\top (\eta + \xi) \end{bmatrix} \\
&= (\eta + \xi)^\top \left(U \Lambda_1 U^\top + V \Lambda_2 V^\top \right) (\eta + \xi) \\
&= (\eta + \xi)^\top \left(\frac{1}{l_1^2} U U^\top + \frac{1}{l_2^2} V V^\top \right) (\eta + \xi) \\
&= \frac{1}{l_1^2} (\eta + \xi)^\top U U^\top (\eta + \xi) + \frac{1}{l_2^2} (\eta + \xi)^\top V V^\top (\eta + \xi) \\
&= \frac{1}{l_1^2} (\eta + \xi)^\top \eta + \frac{1}{l_2^2} (\eta + \xi)^\top \xi \\
&= \frac{\|\eta\|^2}{l_1^2} + \frac{\|\xi\|^2}{l_2^2} \\
&\leq 1,
\end{aligned}$$

which implies

$$\|\eta\| \leq l_1 \quad \text{and} \quad \|\xi\| \leq l_2.$$

Hence,

$$\begin{aligned}
f(x) &= f(x^\star + \eta + \xi) \\
&\leq f(x^\star + \eta) + L_2 \|\xi\| \\
&\leq f(x^\star) + L_1 \|\eta\| + L_2 \|\xi\| \\
&\leq f(x^\star) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\
&= f^\star + \varepsilon
\end{aligned}$$

and so $x \in G_\varepsilon$. Since $x \in \mathcal{E}_E(x^*)$ was arbitrary, this proves that $\mathcal{E}_E(x^*) \subseteq G_\varepsilon$. □

Proposition 3.2.1 lies at the heart of why we should expect, for AD functions, an improved lower bound (at least in theory) on the probability that (RP$\mathcal{X}$) is ε-successful compared to the bound established in [10] under the sole assumption of Lipschitz continuity. Whereas the previous work considered a D-dimensional closed Euclidean ball $B_{\varepsilon/L}(x^*)$ with radius ε/L, where L denotes the Lipschitz constant of the function, the anisotropic structure of AD functions significantly refines this approach. More precisely, with AD functions, we now have access to an ellipsoid whose semi-axis lengths scale inversely with L_1 and L_2. Conversely, the corresponding Euclidean ball

has a radius inversely proportional to $L_1 + L_2$ rather than L_1 or L_2 individually. This can lead to a significantly smaller radius, especially when L_2 is small, compared to the ellipsoid semi-axis of length l_2.

Proposition 3.2.2 There holds

$$\mathrm{Vol}(\mathcal{E}_E(x^*)) = \frac{(L_1 + L_2)^D}{2^D L_1^{d_a} L_2^{D-d_a}} \mathrm{Vol}(B_{\varepsilon/(L_1+L_2)}(x^*)),$$

where $\mathrm{Vol}(C)$ denotes the D-dimensional volume of a set $C \subseteq \mathbb{R}^D$.

Proof. By Theorem 2.1.1, we have

$$\mathcal{E}_E(x^*) = MB_{\varepsilon/(L_1+L_2)}(x^*) + p,$$

where

$$M := \frac{L_1 + L_2}{\varepsilon} Q\Lambda^{-1/2} \quad \text{and} \quad p := x^* - Mx^*.$$

Since volumes are invariant under translations, there holds

$$\mathrm{Vol}(\mathcal{E}_E(x^*)) = \mathrm{Vol}(MB_{\varepsilon/(L_1+L_2)}(x^*)).$$

It is well-known (see [19]) that volumes, under a linear transformation $A \subseteq \mathbb{R}^{D\times D}$, scale with $|\det(A)|$. Hence, we get

$$\begin{aligned}
\mathrm{Vol}(\mathcal{E}_E(x^*)) &= |\det(M)| \cdot \mathrm{Vol}(B_{\varepsilon/(L_1+L_2)}(x^*)) \\
&= \left(\frac{L_1 + L_2}{\varepsilon}\right)^D l_1^{d_a} l_2^{D-d_a} \mathrm{Vol}(B_{\varepsilon/(L_1+L_2)}(x^*)) \\
&= \left(\frac{L_1 + L_2}{\varepsilon}\right)^D \left(\frac{\varepsilon}{2L_1}\right)^{d_a} \left(\frac{\varepsilon}{2L_2}\right)^{D-d_a} \mathrm{Vol}(B_{\varepsilon/(L_1+L_2)}(x^*)) \\
&= \frac{(L_1 + L_2)^D}{2^D L_1^{d_a} L_2^{D-d_a}} \mathrm{Vol}(B_{\varepsilon/(L_1+L_2)}(x^*)).
\end{aligned}$$

□

Corollary 3.2.1 Let $\gamma := L_1/L_2$. If

$$\left(\frac{1+\gamma}{2}\right)^D \geq \gamma^{d_a},$$

then

$$\mathrm{Vol}(\mathcal{E}_E(x^*)) \geq \mathrm{Vol}(B_{\varepsilon/(L_1+L_2)}(x^*)).$$

Proof. By Proposition 3.2.2, it is enough to show that

$$\frac{(L_1 + L_2)^D}{2^D L_1^{d_a} L_2^{D-d_a}} \geq 1.$$

There holds

$$\begin{aligned}\frac{(L_1 + L_2)^D}{2^D L_1^{d_a} L_2^{D-d_a}} &= \frac{(\gamma L_2 + L_2)^D}{2^D (\gamma L_2)^{d_a} L_2^{D-d_a}} \\ &= \frac{(1+\gamma)^D}{2^D \gamma^{d_a}} \\ &\geq 1,\end{aligned}$$

where the last inequality comes from the assumption of the statement. □

Corollary 3.2.2 If

$$d_a \leq \frac{D}{2},$$

then

$$\mathrm{Vol}(\mathcal{E}_E(x^*)) \geq \mathrm{Vol}(B_{\varepsilon/(L_1+L_2)}(x^*)).$$

Proof. By the arithmetic-geometric mean inequality, there holds

$$\frac{1+\gamma}{2} \geq \sqrt{\gamma}$$

and so

$$\left(\frac{1+\gamma}{2}\right)^D \geq \gamma^{D/2}.$$

Since $\gamma \geq 1$ by definition of an AD function and $d_a \leq D/2$ by assumption, we have $\gamma^{D/2} \geq \gamma^{d_a}$ and so

$$\left(\frac{1+\gamma}{2}\right)^D \geq \gamma^{d_a}.$$

The conclusion follows by Corollary 3.2.1. □

The preceding corollaries suggest that, under some conditions on the ratio L_1/L_2 or the anisotropic dimensionality d_a, the volume of $\mathcal{E}_E(x^*)$ exceeds that of the Euclidean ball $B_{\varepsilon/(L_1+L_2)}(x^*)$. Consequently, we can expect a higher probability for (RP$\mathcal{X}$) to be ε-successful when using the ellipsoid instead of the ball, as the former occupies a larger portion of the ambient space than the latter. Note however that this is just a heuristic rather than a formal proof. Moreover, the parameter p has not been taken into account in this reasoning.

3.2.2 Lower bound and convex geometry

We have now everything at our disposal to investigate a lower bound on the probability that (RP$\mathcal{X}$) is ε-successful. For this to occur, the random embedding must intersect the set G_ε to ensure that solving the reduced randomized problem results in an ε-minimizer of the original problem (P). Formally,

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \varepsilon\text{-successful}] = \mathbb{P}[p + \text{range}(A) \cap G_\varepsilon \neq \emptyset]. \tag{3.1}$$

By exploiting the particular structure of the objective function of (P) (and (RP$\mathcal{X}$)), we can derive the following proposition.

Proposition 3.2.3 Let $A \in \mathbb{R}^{D\times d}$ be a Gaussian matrix and $p \in \mathcal{X}$ a given vector. Then,

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \varepsilon\text{-successful}] \geq \mathbb{P}[p + \text{range}(A) \cap \mathcal{E}_E(x^*) \neq \emptyset]. \tag{3.2}$$

Proof. By Proposition 3.2.1, we know that $\mathcal{E}_E(x^*) \subseteq G_\varepsilon$. Hence, we have

$$\{p + \text{range}(A) \cap \mathcal{E}_E(x^*) \neq \emptyset\} \subseteq \{p + \text{range}(A) \cap G_\varepsilon \neq \emptyset\}$$

and the conclusion easily follows from (3.1). □

If $p \in G_\varepsilon$, then clearly (RP$\mathcal{X}$) is ε-successful with probability one, since $p \in p + \text{range}(A)$. So let us assume from now on that $p \in \mathcal{X} \setminus G_\varepsilon$. To proceed, consider the set $C_p(x^*)$ that contains all the rays originating at p and passing through the points in the ellipsoid $\mathcal{E}_E(x^*)$. More precisely,

$$C_p(x^*) := \{p + \theta(x - p) : \theta \geq 0, x \in \mathcal{E}_E(x^*)\}. \tag{3.3}$$

Note that $C_p(x^*)$ is a convex cone that has been translated by p. Specifically, it is the elliptical cone with apex p tangent to $\mathcal{E}_E(x^*)$, as illustrated in Figure 3.2.

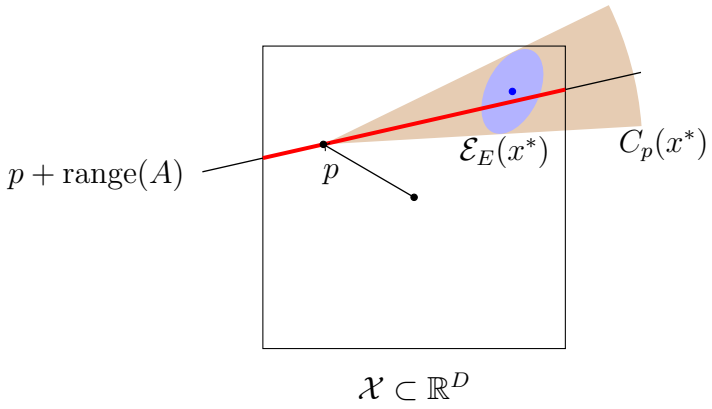


Figure 3.2: (Inspired from [10]) Illustration of (RP$\mathcal{X}$) and the set $C_p(x^*)$. The points along $p + \text{range}(A)$ contained in $\mathcal{X}$ are shown in red.

From the figure above, we immediately see that $p + \text{range}(A)$ intersects $\mathcal{E}_E(x^*)$ if and only if $p + \text{range}(A)$ shares a ray with $C_p(x^*)$. Based on this observation, (3.2) can be expressed as follows.

Theorem 3.2.1 Let $A \in \mathbb{R}^{D\times d}$ be a Gaussian matrix and $p \in \mathcal{X} \setminus G_\varepsilon$. Let $C_p(x^*)$ be defined as in (3.3). Then,

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \varepsilon\text{-successful}] \geq \mathbb{P}[p + \text{range}(A) \cap C_p(x^*) \neq \{p\}]. \tag{3.4}$$

Proof. Same strategy as in [10]. □

Corollary 3.2.3 Let $A \in \mathbb{R}^{D\times d}$ be a Gaussian matrix and $p \in \mathcal{X} \setminus G_\varepsilon$. Then,

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \varepsilon\text{-successful}] \geq \mathbb{P}[\text{range}(A) \cap \mathcal{C}_{\mathcal{E}_E(\Delta)} \neq \{0\}], \tag{3.5}$$

where $\Delta := x^* - p$, $\mathcal{C}_{\mathcal{E}_E(\Delta)}$ is the elliptical cone emanating from the origin, tangent to $\mathcal{E}_E(\Delta)$ and given by

$$\mathcal{C}_{\mathcal{E}_E(\Delta)} := \mathcal{A}\ \text{Circ}_D\left(\alpha^*\right)$$

where

$$\mathcal{A} := Q\Lambda^{-1/2}H, \quad H = I - 2\frac{(e_1 - u)(e_1 - u)^\top}{(e_1 - u)^\top(e_1 - u)}, \quad u = \frac{\Lambda^{1/2}Q^\top\Delta}{\|\Lambda^{1/2}Q^\top\Delta\|}$$

and

$$\alpha^* = \arcsin\left(\frac{1}{\|\Lambda^{1/2}Q^\top\Delta\|}\right).$$

Proof. This is straightforward from Theorem 3.2.1, the fact that $C_p(x^*) - p$ is the elliptical cone emanating from the origin tangent to $\mathcal{E}_E(x^*) - p$ and Theorem 2.1.2. □

When we introduced conic integral geometry, the statistical dimension and the Gaussian width, we saw several ways to quantify the probability that a randomly rotated linear subspace shares a ray with a fixed convex cone (or at least a lower bound). With these tools at hand, we could lower bound the right-hand side of (3.5). However, in our case, we are not directly working with uniformly random rotations of subspaces, but rather with the column space of a Gaussian matrix. The following theorem establishes the link between the two representations.

Theorem 3.2.2 (From [10]) Let $A \in \mathbb{R}^{D\times d}$ be a Gaussian matrix, $Q \in \mathbb{R}^{D\times D}$ a random orthogonal matrix drawn uniformly from the set of all $D \times D$ real orthogonal matrices and $\mathcal{L}_d$ a d-dimensional linear subspace in $\mathbb{R}^D$. Then,

$$\text{range}(A) \overset{law}{=} Q\mathcal{L}_d.$$

Proof. See [10]. □

Corollary 3.2.4 Let $A \in \mathbb{R}^{D\times d}$ be a Gaussian matrix, $Q \in \mathbb{R}^{D\times D}$ a random orthogonal matrix drawn uniformly from the set of all $D \times D$ real orthogonal matrices and $\mathcal{L}_d$ a d-dimensional linear subspace in $\mathbb{R}^D$. Furthermore, let $p \in \mathcal{X} \setminus G_\varepsilon$ and $\mathcal{C}_{\mathcal{E}_E(\Delta)}$ be the elliptical cone tangent to $\mathcal{E}_E(\Delta) := \mathcal{E}_E(x^*) - p$. Then,

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \varepsilon\text{-successful}] \geq \mathbb{P}[Q\mathcal{L}_d \cap \mathcal{C}_{\mathcal{E}_E(\Delta)} \neq \{0\}], \tag{3.6}$$

Proof. Straightforward from Corollary 3.2.3 and Theorem 3.2.2. □

3.2.3 Lower bound using the conic intrinsic volumes

The right-hand side of (3.6) is exactly in the form required to apply the Crofton formula introduced in the conic intrinsic volumes section. This leads to the following theorem.

Theorem 3.2.3 Let $A \in \mathbb{R}^{D\times d}$ be a Gaussian matrix, $p \in \mathcal{X} \setminus G_\varepsilon$ and $\mathcal{C}_{\mathcal{E}_E(\Delta)}$ be the elliptical cone tangent to $\mathcal{E}_E(\Delta) := \mathcal{E}_E(x^*) - p$. Then,

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \varepsilon\text{-successful}] \geq 2h_{D-d+1}, \tag{3.7}$$

with

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

Proof. Just apply Crofton formula to Corollary 3.2.4. □

The lower bound (3.7) is of limited practical use, as we lack effective methods to compute or even estimate the conic intrinsic volumes of a general elliptical cone. Given the limitations of this approach, we focus instead on probability lower bounds derived via the statistical dimension and the Gaussian width.

3.2.4 Lower bound using the statistical dimension

The right-hand side of (3.6) is exactly in the form required to apply the approximate Crofton formula introduced in the statistical dimension section. This leads to the following theorem.

Theorem 3.2.4 Let $A \in \mathbb{R}^{D\times d}$ be a Gaussian matrix, $p \in \mathcal{X} \setminus G_\varepsilon$ and $\mathcal{C}_{\mathcal{E}_E(\Delta)}$ be the elliptical cone tangent to $\mathcal{E}_E(\Delta) := \mathcal{E}_E(x^*) - p$. Then, if

$$\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)}) \geq D - d, \tag{3.8}$$

there holds

$$\mathbb{P}[(\mathrm{RP}\mathcal{X}) \text{ is } \varepsilon\text{-successful}] \geq 1 - 4\exp\left(\frac{-(\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)}) - D + d)^2/8}{\delta_{\min}(\mathcal{C}_{\mathcal{E}_E(\Delta)}) + \delta(\mathcal{C}_{\mathcal{E}_E(\Delta)}) - D + d}\right), \tag{3.9}$$

where

$$\delta_{\min}(C) := \min(\delta(C), \delta(C^*))$$

for a closed convex cone $C \subseteq \mathbb{R}^D$.

Proof. Just apply the approximate Crofton formula to Corollary 3.2.4. □

3.2.5 Lower bound using the Gaussian width

Unlike the previous two cases, the right-hand side of (3.6) does not directly match the form required to apply Gordon's theorem, as presented in the Gaussian width section. In fact, Gordon's theorem provides a lower bound on the probability of non-intersection, whereas we are interested in a lower bound on the probability of intersection. Nevertheless, we can still leverage Gordon's result, thanks to the following theorem, which establishes a relationship between these two quantities.

Theorem 3.2.5 Let X be a linear subspace of $\mathbb{R}^D$ and $P \subseteq \mathbb{R}^D$ be a closed convex cone with $\mathrm{int}(P^*) \neq \emptyset$. There holds

$$X \cap P \neq \{0\} \Leftrightarrow X^\perp \cap \mathrm{int}(P^*) = \emptyset \tag{3.10}$$

where

$$P^* := \{y \in \mathbb{R}^D \mid \langle y, x\rangle \geq 0, \forall x \in P\}$$

is the dual cone of P.

Proof. $\boxed{\Rightarrow}$: Suppose by contradiction that $X^\perp \cap \operatorname{int}(P^*) \neq \emptyset$. Let $x \in X \cap P$ with $x \neq 0$ and $y \in X^\perp \cap \operatorname{int}(P^*)$. By point (ii) of Proposition 2.2.1,

$$\operatorname{int}(P^*) = \{z \in \mathbb{R}^D \mid \langle z, u\rangle > 0, \forall u \in P \setminus \{0\}\}.$$

Therefore,

$$\langle y, x\rangle = 0$$

since $y \in X^\perp$ and $x \in X$ but

$$\langle y, x\rangle > 0$$

since $y \in \operatorname{int}(P^*)$ and $x \in P \setminus \{0\}$, which is a contradiction. Hence, $X^\perp \cap \operatorname{int}(P^*) = \emptyset$.

$\boxed{\Leftarrow}$: Since $X^\perp$ and $\operatorname{int}(P^*)$ are convex, disjoint, nonempty subsets of $\mathbb{R}^D$ and $\operatorname{int}(P^*)$ is open, by the hyperplane separation theorem, there exist a nonzero vector $v \in \mathbb{R}^D$ and a real number $c \in \mathbb{R}$ such that

$$\langle y, v\rangle \leq c < \langle x, v\rangle$$

for all $x \in \operatorname{int}(P^*)$ and $y \in X^\perp$. Since $y \in X^\perp$ and $X^\perp$ is a linear subspace, $\lambda y \in X^\perp$ for all $\lambda > 0$ and so

$$\langle y, v\rangle \leq \frac{c}{\lambda}.$$

Similarly, $-\lambda y \in X^\perp$ and so

$$-\frac{c}{\lambda} \leq \langle y, v\rangle.$$

Hence,

$$-\frac{c}{\lambda} \leq \langle y, v\rangle \leq \frac{c}{\lambda}$$

and letting $\lambda \to +\infty$ gives

$$\langle y, v\rangle = 0.$$

Since this is true for all $y \in X^\perp$, there holds $v \in (X^\perp)^\perp = X$. Furthermore, this means that

$$0 \leq c < \langle x, v\rangle$$

for all $x \in \text{int}(P^*)$. Note that by point (i) of Proposition 2.2.1, P^* is closed and convex. We know that if C is a convex set with nonempty interior, then the closure of C is equal to the closure of the interior of C. In our case, this means that

$$\overline{\text{int}(P^*)} = \overline{P^*} = P^*$$

where the last equality comes from the closedness of P^*. Let $x \in P^*$. By the properties of the closure of a set, there exists a sequence $(x_n) \subseteq \text{int}(P^*)$ such that $x_n \to x$. Since strong convergence implies weak convergence, there holds

$$\langle x, v \rangle = \lim_{n \to +\infty} \langle x_n, v \rangle \geq 0.$$

Since this is true for all $x \in P^*$, this means that $v \in (P^*)^*$. But by point (iii) of Proposition 2.2.1, since P is a closed convex cone,

$$(P^*)^* = P$$

and so $v \in P$. As $v \neq 0$, it follows that $X \cap P \neq \{0\}$ as expected. □

Therefore, using Theorem 3.2.5 (assuming its condition), we deduce that

$$\begin{aligned}
\mathbb{P}[Q\mathcal{L}_d \cap \mathcal{C}_{\mathcal{E}_E(\Delta)} \neq \{0\}] &= \mathbb{P}[(Q\mathcal{L}_d)^\perp \cap \text{int}(\mathcal{C}^*_{\mathcal{E}_E(\Delta)}) = \emptyset] && (3.11)\\
&= \mathbb{P}[(Q\mathcal{L}_d)^\perp \cap \text{int}(\mathcal{C}^*_{\mathcal{E}_E(\Delta)}) \cap S^{D-1} = \emptyset] && (3.12)\\
&= \mathbb{P}[(Q\mathcal{L}_d)^\perp \cap (\mathcal{C}^*_{\mathcal{E}_E(\Delta)} \cap S^{D-1}) = \emptyset], && (3.13)
\end{aligned}$$

where (3.12) comes from the scale invariance of cones and (3.13) from the fact that a set with smooth boundary has measure 0. We are now ready to state the following theorem.

Theorem 3.2.6 Let $A \in \mathbb{R}^{D \times d}$ be a Gaussian matrix, $p \in \mathcal{X} \setminus G_\varepsilon$ and $\mathcal{C}_{\mathcal{E}_E(\Delta)}$ be the elliptical cone tangent to $\mathcal{E}_E(\Delta) := \mathcal{E}_E(x^*) - p$. If

$$w(\mathcal{C}^*_{\mathcal{E}_E(\Delta)} \cap S^{D-1}) < \sqrt{d}, \tag{3.14}$$

then

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \varepsilon\text{-successful}] \geq 1 - \frac{5}{2} \exp\left(- \frac{(d/\sqrt{d+1} - w(\mathcal{C}^*_{\mathcal{E}_E(\Delta)} \cap S^{D-1}))^2}{18} \right) \tag{3.15}$$

Proof. $(Q\mathcal{L}_d)^\perp$ is a $(D-d)$-dimensional linear subspace and by point (i) of Proposition 2.2.1, $\mathcal{C}^*_{\mathcal{E}_E(\Delta)}$ is a closed convex cone. Therefore, $\mathcal{C}^*_{\mathcal{E}_E(\Delta)} \cap S^{D-1}$ is a closed subset of the unit sphere and Gordon's theorem can be applied to (3.13). □

3.3 Analysis of the lower bounds

3.3.1 Using the statistical dimension

In Theorem 3.2.4, we established a lower bound on the probability that (RP$\mathcal{X}$) is ε-successful, based on the statistical dimension of the elliptical cone tangent to $\mathcal{E}_E(\Delta)$, the ellipsoid of ε-minimizers translated by p. For this lower bound to be useful, we need a way to compute this statistical dimension, at least numerically. Fortunately, as explained under Theorem 2.3.2, this is possible. Specifically, we have

$$\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)}) = \mathbb{E}\left[\|\Pi_{\mathcal{C}_{\mathcal{E}_E(\Delta)}}(g)\|^2\right],$$

where $g \sim \mathcal{N}(0, I_n)$. The main computational challenge lies in evaluating the projection of g onto the elliptical cone. Unfortunately, there does not seem to exist a closed-form expression for this projection. However, in the case of a circular cone, such an expression is known (cf. [35]). Since

$$\mathcal{C}_{\mathcal{E}_E(\Delta)} = \mathcal{A}\mathrm{Circ}_D(\alpha^*) \quad \text{where} \quad \mathcal{A} := (Q\Lambda^{-1/2}H),$$

the original projection problem

$$\Pi_{\mathcal{C}_{\mathcal{E}_E(\Delta)}}(g) := \underset{y \in \mathcal{C}_{\mathcal{E}_E(\Delta)}}{\arg\min} \|y - g\|^2$$

can be reformulated as

$$\Pi_{\mathcal{C}_{\mathcal{E}_E(\Delta)}}(g) = \mathcal{A} \underset{x \in \mathrm{Circ}_D(\alpha^*)}{\arg\min} \|\mathcal{A}x - g\|^2,$$

where now the set over which we are optimizing is much simpler (it is a circular cone). Based on this idea, we decided to implement the Nesterov's accelerated projected gradient method to solve the projection problem efficiently. Indeed,

- we have a smooth, convex function with Lipschitz continuous gradient (and known Lipschitz constant L):

  $$\nabla\left(\|\mathcal{A}x - g\|^2\right) = 2\mathcal{A}^\top(\mathcal{A}x - g), \quad \nabla^2\left(\|\mathcal{A}x - g\|^2\right) = 2\mathcal{A}^\top\mathcal{A}$$

  which means $L = 2\|\mathcal{A}\|^2 = 2l_2^2$,

- we have an efficient way of computing the projection onto the feasible set (the closed-form expression for the projection onto a circular cone) and,

- the accelerated projected gradient method benefits from an optimal convergence rate for first order methods ($\mathcal{O}(1/k^2)$).

The implementation was carried out in C++ for improved computational performance and parallelized to efficiently estimate the expectation, which is just an average of squared norms of solutions to independent optimization problems. With this code in place, we are now fully equipped to begin the analysis of the lower bound based on the statistical dimension of $\mathcal{C}_{\mathcal{E}_E(\Delta)}$.

We first focus, in a theoretical manner, on the condition (3.8), which can be rewritten as

$$\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)}) + d \geq D. \tag{3.16}$$

This inequality is a necessary condition for Theorem 3.2.4 to ensure that the elliptical cone is most likely to share a ray with a randomly rotated linear subspace of dimension d. Therefore, to ensure that (RP$\mathcal{X}$) will be ε-successful with high probability, the sum of the statistical dimension of the elliptical cone and the dimension of the random embedding dimension in (RP$\mathcal{X}$) should exceed the dimension of the ambient space. Note that this result is also consistent by dimensionality analysis, as the statistical dimension serves as a natural extension of the notion of dimension of linear subspaces to closed convex cones. Accordingly, all the terms in the bound above can be viewed as dimensions. We emphasize that this result was already known for linear subspaces, as stated below.

Proposition 3.3.1 Let $X \subseteq \mathbb{R}^D$ be a m-dimensional linear subspace and $Y \subseteq \mathbb{R}^D$ be a n-dimensional linear subspace. If

$$m + n > D,$$

then

$$X \cap Y \neq \{0\}.$$

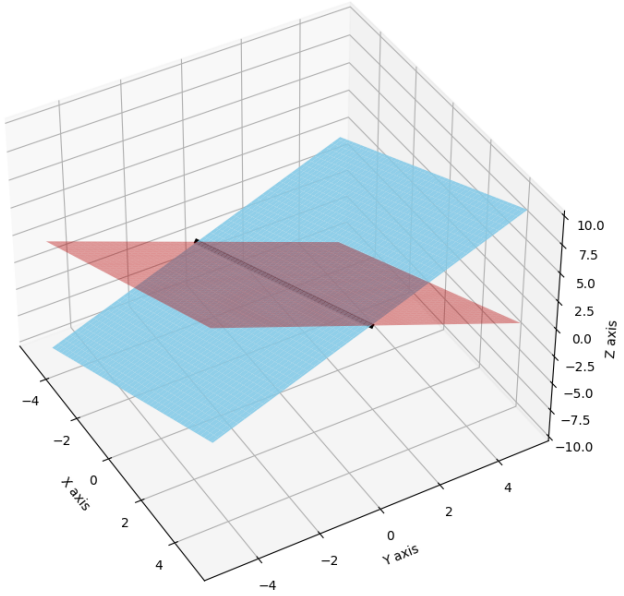


Figure 3.3: Illustration of Proposition 3.3.1 with $m = n = 2$ and $D = 3$. Since $m + n > D$, the intersection of the planes is non-trivial. Here, this intersection is a line, shown in black.

To prove Proposition 3.3.1, we need the following lemma.

Lemma 3.3.1 (Grassmann formula for linear subspaces of $\mathbb{R}^D$) Let $X, Y \subseteq \mathbb{R}^D$ be linear subspaces. There holds

$$\dim(X \cap Y) = \dim(X) + \dim(Y) - \dim(X + Y).$$

Proof. (Proposition 3.3.1) Since the sum of linear subspaces of $\mathbb{R}^D$ lies in $\mathbb{R}^D$, we have

$$\dim(X + Y) \leq D$$

and so, by Lemma 3.3.1, there holds

$$\dim(X \cap Y) = \dim(X) + \dim(Y) - \dim(X + Y) \geq m + n - D > 0.$$

Hence, X has a non-trivial intersection with Y, which is what we wanted to prove. □

We now turn our attention to the case where the objective function in (P) varies infinitely slowly along the orthogonal subspace $\mathcal{T}^\perp$ (which means that the function remains constant along $\mathcal{T}^\perp$). More precisely, we consider AD functions with $L_2 = 0$. Such functions are already known in the literature as *functions with (low) effective dimensionality*.

Definition 3.3.1 (Taken from [10]) A function $f : \mathbb{R}^D \to \mathbb{R}$ is said to have effective dimensionality d_e, with $d_e \leq D$, if

- there exists a linear subspace $\mathcal{T}$ of dimension d_e such that for all $x_\top \in \mathcal{T} \subseteq \mathbb{R}^D$ and $x_\perp \in \mathcal{T}^\perp$, we have
$$f(x_\top + x_\perp) = f(x_\top),$$
- d_e is the smallest integer with this property.

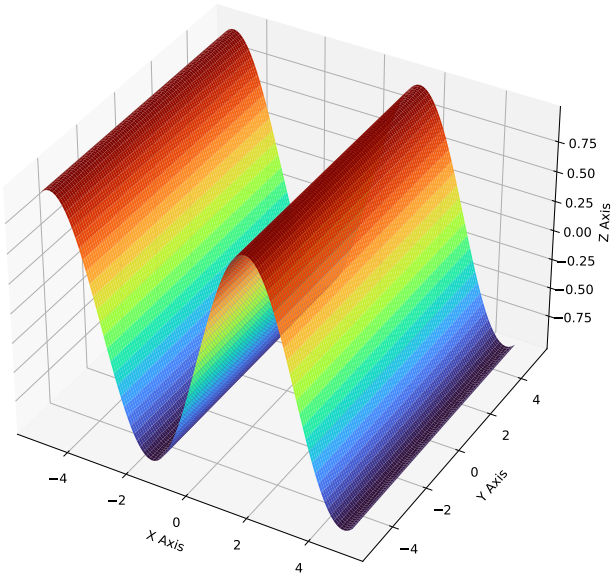


Figure 3.4: Example of function with effective dimensionality $d_e = 1$. Here, $f(x, y) = \sin(x)$.

This class of functions, while lacking theoretical guarantees for small effective dimensions d_e, seems to arise in the objectives of highly overparameterized optimization problems, such as those encountered in deep neural network training. In fact, in such settings, one might expect that not all parameters in a neural network (especially in high dimensions) have an impact on the training loss function of the network. The hope is that, with a high number of parameters, the training loss only varies along the subspace formed by the range of a small number of them. Hence, instead of solving the higher-dimensional problem (P), we could solve (a sequence of) the lower-dimensional problem (RP$\mathcal{X}$). A very useful result regarding this type of functions is presented below.

Theorem 3.3.1 (Adapted from [33]) Let $f : \mathbb{R}^D \to \mathbb{R}$ be a function with effective dimensionality d_e, $A \in \mathbb{R}^{D\times d}$ be a Gaussian matrix and $p \in \mathbb{R}^D$. If

$$d \geq d_e,$$

then with probability 1, for any $x \in \mathbb{R}^D$, there exists a $y \in \mathbb{R}^d$ such that

$$f(x) = f(Ay + p).$$

Proof. See [33]. □

Therefore, for functions with effective dimensionality d_e, solving a single randomized reduced problem (RP$\mathcal{X}$) with $d \geq d_e$ is enough to solve (P) in the unconstrained case $\mathcal{X} = \mathbb{R}^D$. Note that if $\mathcal{X} \neq \mathbb{R}^D$, this is no longer true since Theorem 3.3.1 only guarantees the existence of a vector $y \in \mathbb{R}^d$ with $Ay + p \in \mathbb{R}^D$ (and not $Ay + p \in \mathcal{X}$).

We now demonstrate how our earlier findings based on the statistical dimension align with this result. For this, recall that the ellipsoid $\mathcal{E}_E(\Delta)$ has $D - d_a$ semi-axes whose lengths are inversely proportional to L_2. As $L_2 \to 0$, $\mathcal{E}_E(\Delta)$ degenerates into an infinite hypercylinder, with a circular base of dimension d_a and radius $l_1 = \varepsilon/(2L_1)$, as illustrated in Figure 3.5.

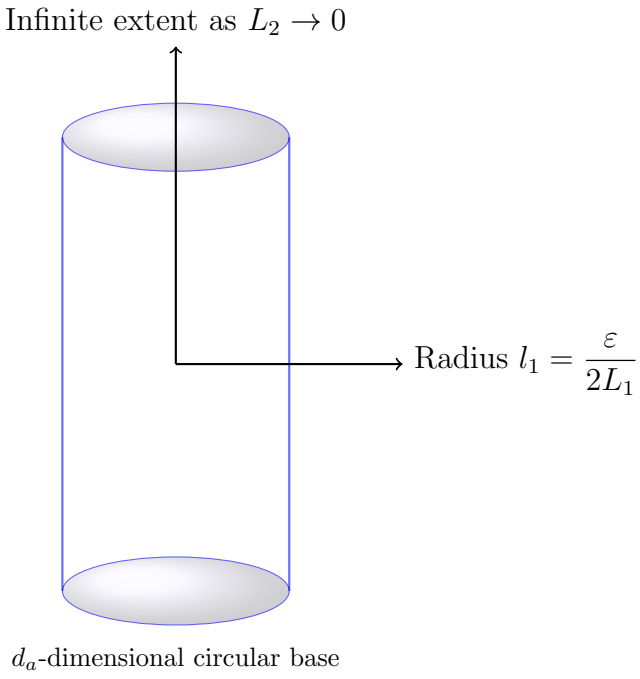


Figure 3.5: Illustration of the ellipsoid $\mathcal{E}_E(\Delta)$ when $L_2 \to 0$.

Hence, we can expect the elliptical cone $\mathcal{C}_{\mathcal{E}_E(\Delta)}$ tangent to $\mathcal{E}_E(\Delta)$ to approach a $(D-d_a)$-dimensional linear subspace of $\mathbb{R}^D$. As the statistical dimension of a subspace is just its dimension, we have

$$\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)}) \approx D - d_a$$

and so, using (3.16), we conclude that $d \gtrsim d_a$ is necessary for Theorem 3.2.4 to ensure that (RP$\mathcal{X}$) is ε-successful with high probability. This agrees with Theorem 3.3.1.

3.3.1.1 Influence of the ratio

In what follows, if not specified, the values of the different parameters used in the numerical experiments are shown in Table 3.1

D	d_a	ε	L_1	$\|x^* - p\|$	γ
100	20	0.1	0.1	10	20

Table 3.1: Values used for the numerical experiments

We demonstrate numerically that the statistical dimension of the elliptical cone $\mathcal{C}_{\mathcal{E}_E(\Delta)}$ becomes close to $D - d_a$ as the ratio $\gamma = L_1/L_2$ tends to infinity (as $L_2 \to 0$) by examining its behavior in the following figure.

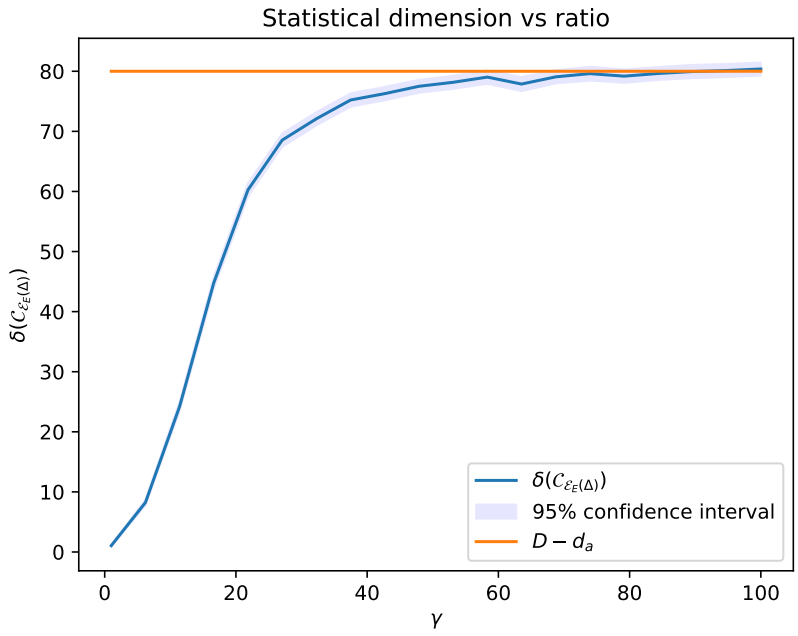


Figure 3.6: Influence of L_1/L_2 on the statistical dimension

As explained earlier, for a large ratio γ, we observe that $\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)}) \approx D - d_a$. However, there exists a regime where this approximation breaks down. More precisely, if p is very close to $\mathcal{E}_E(\Delta)$, we can expect the statistical dimension to exceed $D - d_a$, possibly by a significant margin. In fact, in this scenario, the elliptical cone tangent to $\mathcal{E}_E(\Delta)$ will exhibit a wide opening angle in all directions, causing the d_a semi-axes whose lengths are inversely proportional to L_1 to have more impact on its statistical dimension. The contribution from these axes can be as large as d_a, meaning that the statistical dimension can go up to $D - d_a + d_a = D$. This behavior is illustrated in the figure below.

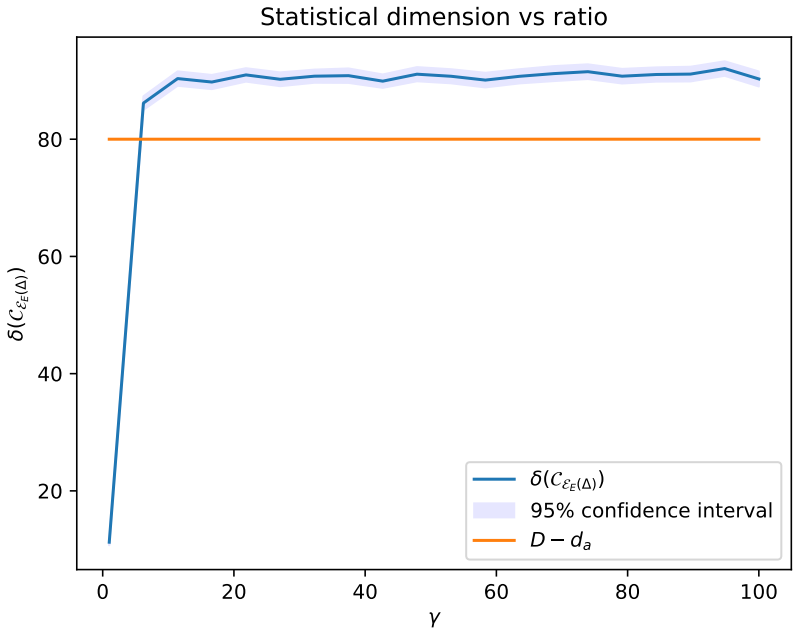


Figure 3.7: The approximation $\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)}) \approx D - d_a$ becomes false when p is very close to $\mathcal{E}_E(\Delta)$. Here, $\|x^* - p\| = 3l_1$.

Fortunately, our main interest lies in the regime where p is not too close to $\mathcal{E}_E(\Delta)$. Moreover, even outside this regime, the approximation $D - d_a$ still serves as a lower bound, although it can be loose.

To continue, we observe that the statistical dimension tends to increase as the ratio γ grows. This behavior makes sense since, when $L_2 \to 0$, the ellipsoid $\mathcal{E}_E(\Delta)$ expands along the orthogonal subspace $\mathcal{T}^\perp$, occupying more and more space. Consequently, the elliptical cone tangent to it becomes wider, which explains the increase in the statistical dimension.

We now examine how the ratio γ affects the probability lower bound (3.9).

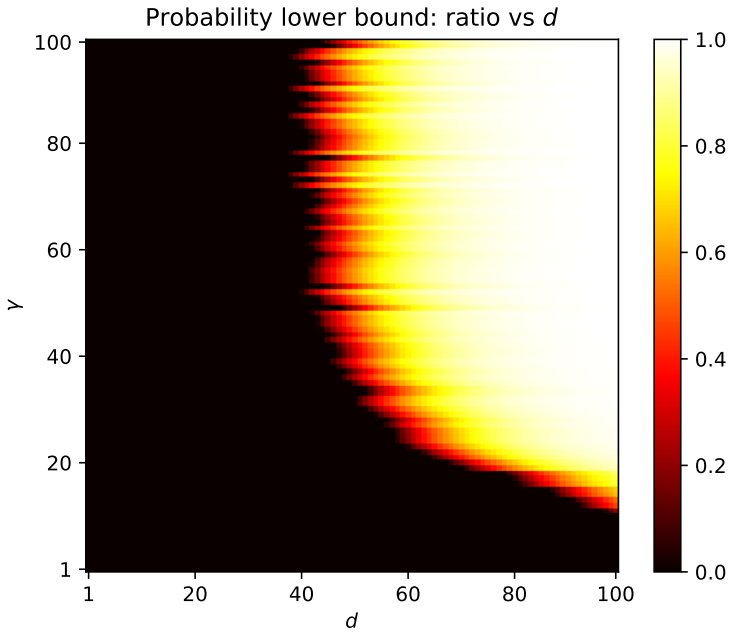


Figure 3.8: Influence of L_1/L_2 on the probability lower bound in Theorem 3.2.4

Looking at Figure 3.8, we notice that for small values of the ratio γ, a larger embedding dimension d is required for (RP$\mathcal{X}$) to be ε-successful with high probability, and vice versa. Specifically, for high values of the ratio, there appears to be a threshold dimension d^* such that the probability of (RP$\mathcal{X}$) being ε-successful primarily depends on whether d exceeds this threshold or not. This observation is in line with our previous analysis regarding functions with effective dimensionality, for which a random embedding of dimension $d \geq d_e$ intersects $\mathcal{E}_E(\Delta)$ with probability one. Note that the observed threshold d^* here seems to exceed d_a. This is because our lower bound (3.9) is based on the approximate Crofton formula from Theorem 2.3.4 rather than the exact form seen in the conic intrinsic volumes section. As a result, we obtain a looser threshold $d^* > d_a$.

3.3.1.2 Influence of the accuracy

In the previous section, we examined the influence of the ratio L_1/L_2 on the statistical dimension and the probability lower bound (3.9). More precisely, we showed numerically that the statistical dimension increases as $L_2 \to 0$. We now turn our attention to the effects of the accuracy ε on this statistical dimension. Since the lengths of the semi-axes of $\mathcal{E}_E(\Delta)$ scale proportionally with ε, we expect the statistical dimension to grow as ε increases. This is indeed illustrated in Figure 3.9.

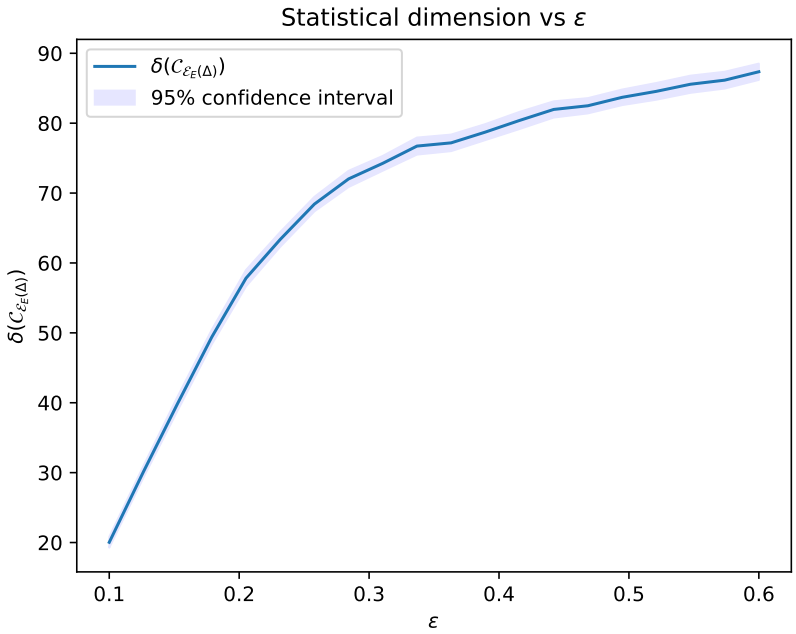


Figure 3.9: Influence of ε on the statistical dimension

Based on the figure above, we also expect the probability lower bound in Theorem 3.2.4 to increase with ε.

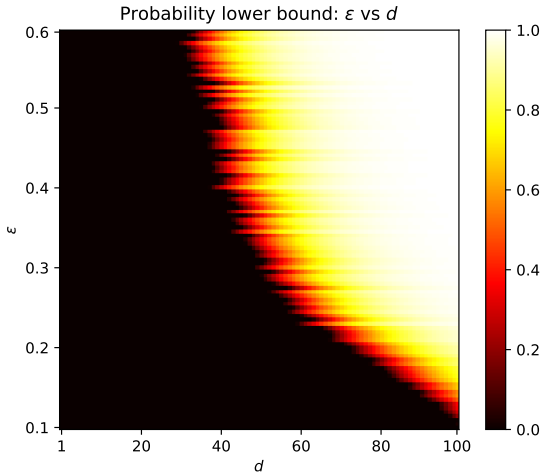


Figure 3.10: Influence of ε on the probability lower bound in Theorem 3.2.4

3.3.1.3 Influence of the anistropic dimension

By definition, a function with anisotropic dimensionality d_a varies slowly along a $(D - d_a)$-dimensional linear subspace and "normally (meaning it is Lipschitz continuous

with a not necessarily small Lipschitz constant) along a d_a-dimensional subspace. Theoretically, it would be better for d_a to be much smaller than the ambient dimension D, as this allows us to exploit the fact that the function primarily varies in only a few directions, thereby simplifying the optimization problem (P).

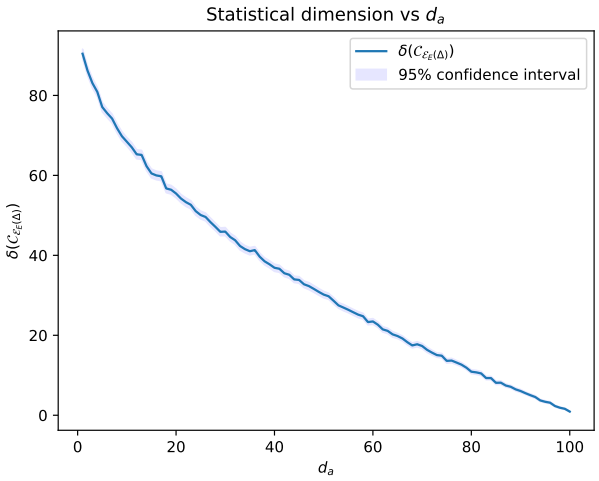


Figure 3.11: Influence of d_a on the statistical dimension.

From Figure 3.11, we observe that the statistical dimension tends to decrease as d_a increases. This is an intuitive result. Indeed, recall that $\mathcal{E}_E(\Delta)$ possesses d_a semi-axes whose lengths are $\varepsilon/(2L_1)$ and $D - d_a$ semi-axes whose lengths are $\varepsilon/(2L_2)$. Since for an AD function, $L_1 \gg L_2$, the semi-axes corresponding to the $D - d_a$ directions are typically much longer. Increasing d_a will then decrease the number of those axes and so the ellipsoid $\mathcal{E}_E(\Delta)$ will occupy a smaller volume in $\mathbb{R}^D$. Consequently, the elliptical cone $\mathcal{C}_{\mathcal{E}_E(\Delta)}$ will become narrower, explaining the decrease in its statistical dimension.

Hence, based on our previous findings, we can expect the probability that (RP$\mathcal{X}$) is ε-successful to decrease with d_a, which is indeed what we observe in Figure 3.12

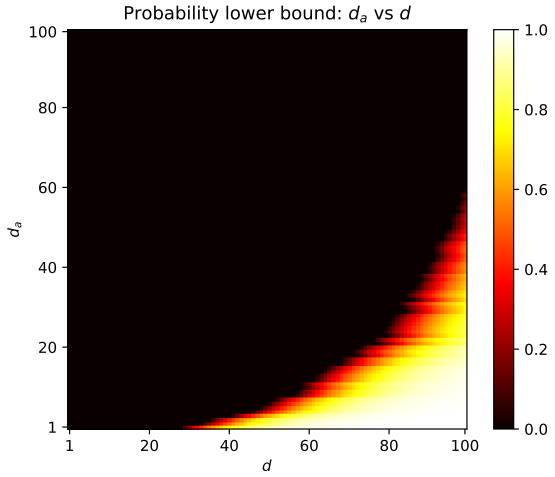


Figure 3.12: Influence of d_a on the probability lower bound in Theorem 3.2.4.

3.3.1.4 Influence of the distance to the minimizer

To continue our analysis of the lower bound using the statistical dimension, we look at the influence of the parameter p on the statistical dimension $\mathcal{C}_{\mathcal{E}_E(\Delta)}$ and probability lower bound (3.9). Recall that $\mathcal{E}_E(\Delta)$ is an ellipsoid centered at $\Delta = x^* - p$ and $\mathcal{C}_{\mathcal{E}_E(\Delta)}$ is the elliptical cone emanating from the origin that is tangent to it.

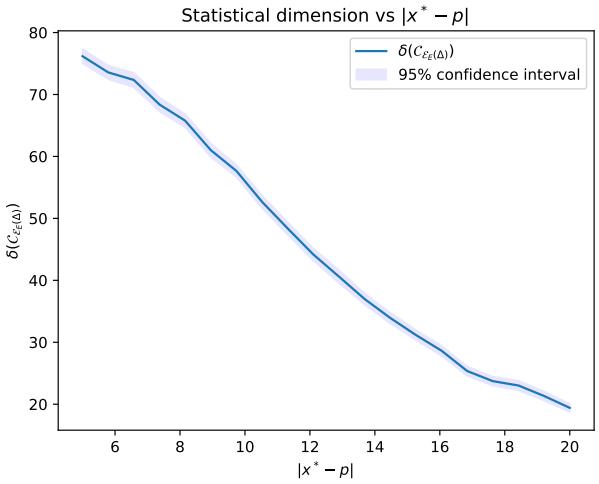


Figure 3.13: Influence of $\|x^* - p\|$ on the statistical dimension

In Figure 3.13, we see that $\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)})$ tends to decrease as $\|x^* - p\|$ increases. The reason behind this behavior is simple. If p is far from x^*, then $x^* - p$ will be far from the origin, which means that $\mathcal{E}_E(\Delta)$ will be far from the origin as well. As a result,

the elliptical cone $\mathcal{C}_{\mathcal{E}_E(\Delta)}$ will be narrow and so its statistical dimension will be small. Consequently, one can expect the probability that (RP$\mathcal{X}$) is ε-successful to decrease with increasing $\|x^* - p\|$, which is what we observe in Figure 3.14.

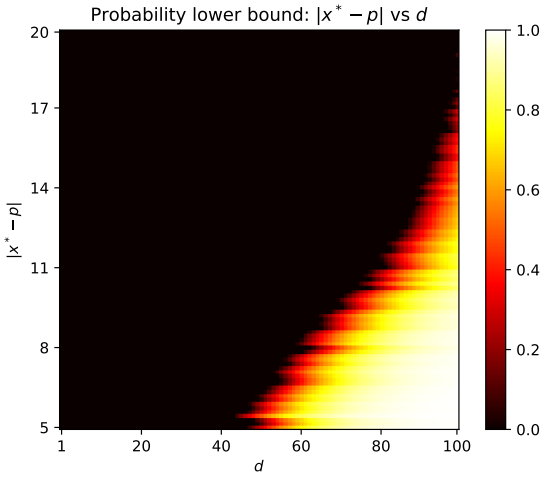


Figure 3.14: Influence of $\|x^* - p\|$ on the probability lower bound in Theorem 3.2.4

3.3.1.5 On a generalization of Theorem 3.3.1

A closer look at Figure 3.6 reveals that the ratio γ does not need to be infinite for the approximation $\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)}) \approx D - d_a$ to hold accurately. For instance, in this example, the minimum ratio such that the approximation becomes good is around $\gamma_{\min} \approx 30$. It would be valuable to establish a relationship between this minimal ratio and the various parameters of (RP$\mathcal{X}$). In fact, when $\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)}) \approx D - d_a$, we have seen, thanks to (3.8), that a single random embedding of sufficiently high dimension $d \geq d_a$ is probably enough for (RP$\mathcal{X}$) to be ε-successful. Based on numerical evidence, we strongly believe that this minimal ratio satisfies $\gamma_{\min} = \mathcal{O}(\Delta/l_1)$.

Conjecture 3.3.1 Let $\Delta := \|x^* - p\|$ be the distance of p to the global minimizer x^* in Assumption 3.2.2 and $l_1 := \varepsilon/(2L_1)$ be the length of each of the d_a semi-axes of $\mathcal{E}_E(\Delta)$ inversely proportional to L_1. Then, the minimal ratio L_1/L_2 such that $\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)}) \approx D - d_a$ is given by

$$\gamma_{\min} = \mathcal{O}\left(\frac{\Delta}{l_1}\right).$$

With the help of Theorem 3.2.4, Conjecture 3.3.1 leads to a generalization of Theorem 3.3.1.

Corollary 3.3.1 Let the function f in (RP$\mathcal{X}$) be a function with anisotropic dimensionality d_a. Suppose that

$$\gamma \geq \gamma_{\min} := \mathcal{O}\left(\frac{\Delta}{l_1}\right). \tag{3.17}$$

If

$$d \gtrsim d_a,$$

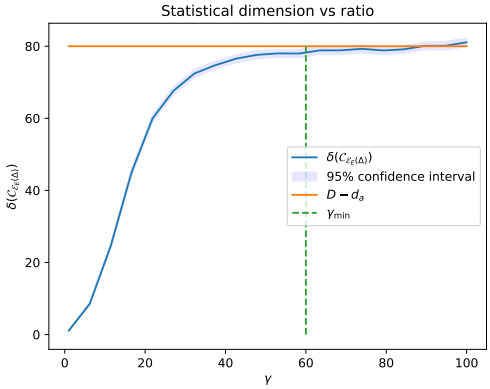
then

$$\mathbb{P}[(\text{RP}\mathcal{X}) \text{ is } \varepsilon\text{-successful}] \approx 1.$$

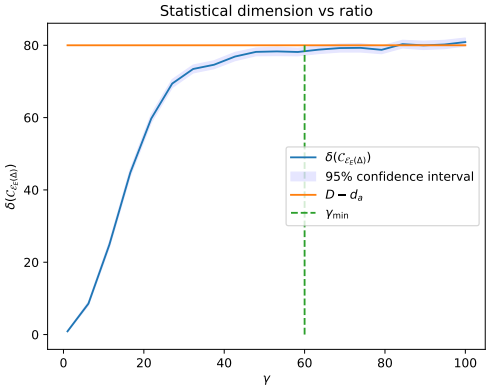
Proof. (Assuming Conjecture 3.3.1) If $\gamma \geq \gamma_{\min}$, we have $\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)}) \approx D - d_a$ and so the condition (3.8) becomes $d \gtrsim d_a$. By Theorem 3.2.4, (RP$\mathcal{X}$) will then most likely be ε-successful. □

In a theoretical point of view, Conjecture 3.3.1 makes a lot of sense. In fact, if $L_2 \to 0$ (meaning that the function f becomes a function with effective dimensionality), then $\gamma \to \infty$ and so we see by Corollary 3.3.1 that, regardless of the distance of p to x^* or the value of ε, (3.17) will be satisfied. Hence, if $d \gtrsim d_a$, (RP$\mathcal{X}$) will be ε-successful with high probability, which agrees with Theorem 3.3.1. Furthermore, by dimensionality analysis, (3.17) seems to be valid. For example, if we choose a unit of distance, let's say meters, then both Δ (a distance) and l_1 (a length) are expressed in meters. Hence, the ratio Δ/l_1 becomes dimensionless, just like the ratio $\gamma := L_1/L_2$. While this reasoning does not constitute a formal proof of Conjecture 3.3.1, it nonetheless provides us a way to convince ourselves of its validity.

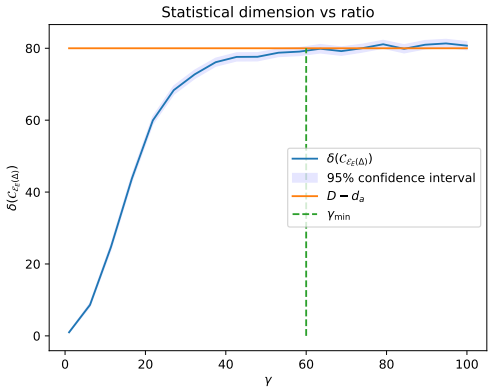
We now validate Conjecture 3.3.1 numerically, by varying Δ and l_1 in such a way that $\gamma_{\min}$ remains constant.



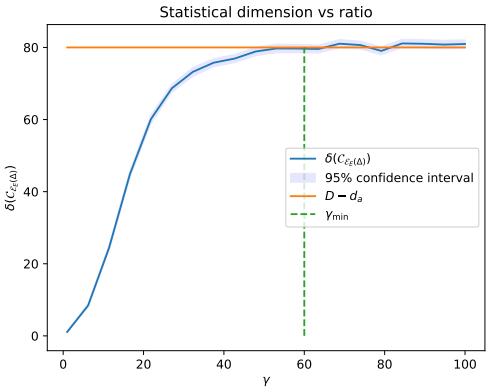
(a) Factor $f = 1$



(b) Factor $f = 5$



(c) Factor $f = 10$



(d) Factor $f = 20$

Figure 3.15: Varying Δ and l_1 in such a way that $\gamma_{\min}$ remains constant ($\gamma_{\min} \approx 60$). For this, we multiply Δ and l_1 by a factor f. It is clear that the approximation $\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)}) \approx D - d_a$ holds well when $\gamma \geq \gamma_{\min}$.

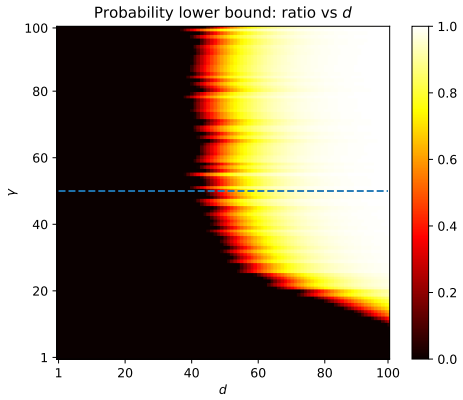


Figure 3.16: Probability lower bound of Theorem 3.2.4. The dashed line indicates the value of $\gamma_{\min}$. Notice that beyond this threshold, the phase transition becomes nearly vertical, meaning that any randomized reduced problem (RP$\mathcal{X}$) with embedding dimension d exceed a critical value d^* will be ε-successful with high probability.

3.3.1.6 Comparison with bound (1.1)

To conclude this section, we demonstrate briefly how our bound based on the statistical dimension in Theorem 3.2.4 improves upon the bound (1.1), except in a few specific cases. For this comparison to be meaningful, recall from Proposition 3.1.1 that AD functions are Lipschitz continuous functions with Lipschitz constant equal to the sum $L_1 + L_2$. Since the derivation of the bound (1.1) in [10] only assumes Lipschitz continuity and uses a ball $B_{\varepsilon/L}(x^*)$ of ε-minimizers, where L is the Lipschitz constant of the objective. Hence, in our setting, this ball becomes $B_{\varepsilon/(L_1+L_2)}(\Delta)$ and we can now compare the two lower bounds.

Figure 3.17 shows the statistical dimension of both $\mathcal{C}_{\mathcal{E}_E(\Delta)}$ and $\mathcal{C}_{B_{\varepsilon/(L_1+L_2)}(\Delta)}$, with respect to the ratio γ.

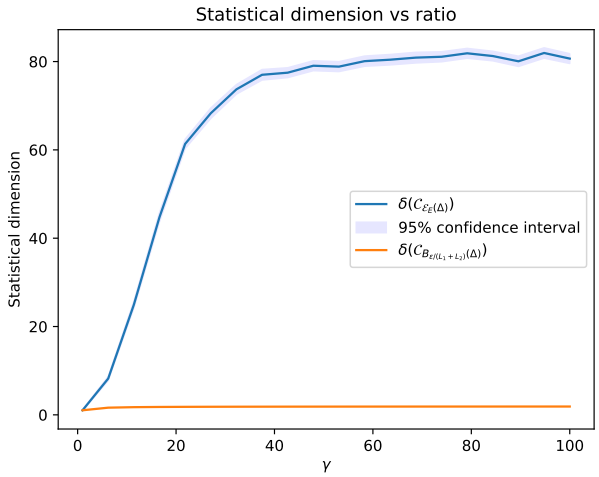
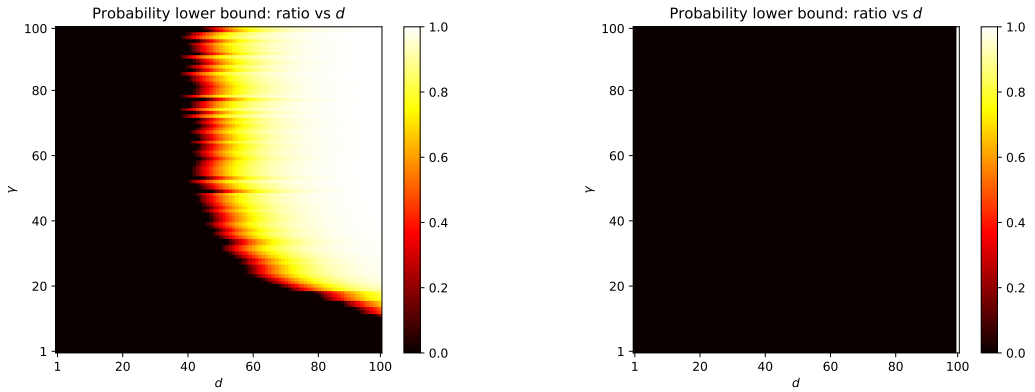


Figure 3.17: Comparison of $\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)})$ and $\delta(\mathcal{C}_{B_{\varepsilon/(L_1+L_2)}(\Delta)})$

We see in the figure above that $\mathcal{C}_{B_{\varepsilon/(L_1+L_2)}(\Delta)}$ doesn't seem to increase a lot with respect to γ. This is because the ball $B_{\varepsilon/(L_1+L_2)}(\Delta)$ doesn't really capture the information provided by functions with anisotropic dimensionality, especially when d_a is small compared to D. Hence, because $\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)})$ is generally much greater than $\delta(\mathcal{C}_{B_{\varepsilon/(L_1+L_2)}(\Delta)})$, we should expect Theorem 3.2.4 to provide a much better bound than (1.1) for functions with anisotropic dimensionality, as illustrated in the figure below.



(a) Lower bound using the elliptical cone (b) Lower bound using the circular cone

Figure 3.18: A comparison of the lower bound of Theorem 3.2.4 and (1.1).

3.3.2 Using the Gaussian width

Following our approach with the statistical dimension bound, we next analyze the bound derived using the Gaussian width. Consider condition (3.14). A consequence of Proposition 2.4.4 and the behavior of the statistical dimension under duality is

$$D - 2 \leq w(\mathcal{C}_{\mathcal{E}_E(\Delta)} \cap S^{D-1})^2 + w(\mathcal{C}^*_{\mathcal{E}_E(\Delta)} \cap S^{D-1})^2 \leq D$$

. Therefore, for large D, we can assume

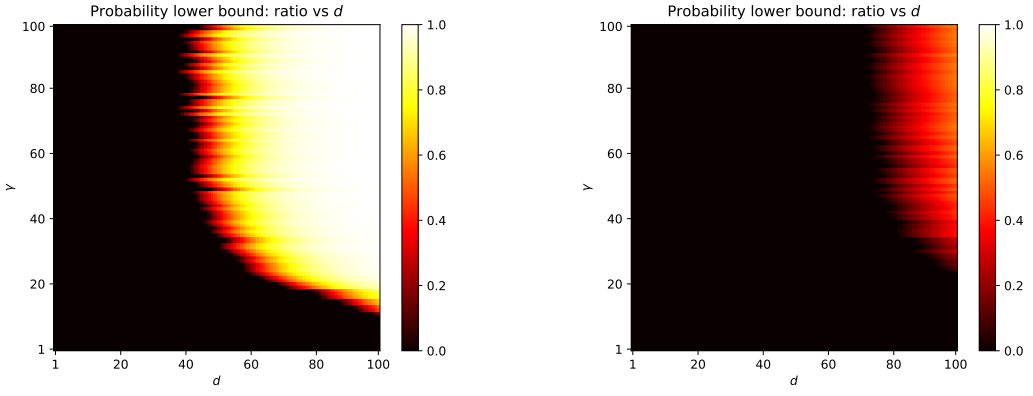
$$w(\mathcal{C}_{\mathcal{E}_E(\Delta)} \cap S^{D-1})^2 + w(\mathcal{C}^*_{\mathcal{E}_E(\Delta)} \cap S^{D-1})^2 \approx D.$$

Hence, condition (3.14) becomes

$$w(\mathcal{C}_{\mathcal{E}_E(\Delta)} \cap S^{D-1})^2 + d > \mathcal{O}(D)$$

as $D \to \infty$. This aligns with (3.16), recalling that, by Proposition 2.4.4, the Gaussian width squared can be interpreted as a dimension.

We could repeat our numerical analysis performed for the bound using the statistical dimension with the bound based on the Gaussian width, and we would observe the same overall results. This similarity stems from the fact that their general forms are similar to each other, mainly differing in their coefficients. Nevertheless, the latter bound appears to yield less sharper results compared to (3.9), as illustrated in the figure below.



(a) Lower bound using the statistical dimension

(b) Lower bound using the Gaussian width

Figure 3.19: A comparison of the probability lower bounds of Theorem 3.2.4 and Theorem 3.2.6. We observe similar overall results, although the latter bound seems to be less sharp.

Given this observation, we chose not to pursue a deeper analysis of the lower bound based on the Gaussian width, as the statistical dimension bound already offers insightful and sharper results.

3.4 Applications

In this section, we explore how AD functions naturally appear in various branches of mathematics, motivating their study.

3.4.1 Parameter optimization in machine learning

We are given a dataset

$$\mathcal{D} := \{(x_i, y_i)\}_{i=1}^n$$

where $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}^K$ and $n, d, K \in \mathbb{N}$. In machine learning, we are often interested in fitting a model

$$f_\theta : \mathbb{R}^d \to \mathbb{R}^K,$$

parameterized by $\theta \in \mathbb{R}^D$ to the dataset $\mathcal{D}$, with $D \in \mathbb{N}$ possibly very large. This means that we want to find the best parameters θ^* such that

$$f_{\theta^*}(x_i) \approx y_i$$

for all $i \in \{1, \ldots, n\}$, in the hope that the model will give good results for unseen new data. One way to achieve this is to consider a training set loss function

$$\mathcal{L} : \mathbb{R}^D \to \mathbb{R}_+$$

defined for parameters $\theta \in \mathbb{R}^D$ by

$$\mathcal{L}(\theta) := \frac{1}{n} \sum_{i=1}^n l(\theta, x_i, y_i)$$

where

$$l : \mathbb{R}^{D \times d \times K} \to \mathbb{R}_+$$

is a loss function defined on each data point (x_i, y_i) and parameters θ. Examples of such functions include the mean squared error

$$\text{MSE}(\theta, x, y) := \|f_\theta(x) - y\|^2$$

and the cross-entropy loss

$$\mathrm{CE}(\theta, x, y) := -\sum_{i=1}^{K} y^{(i)} \log(f_\theta(x)^{(i)}),$$

where $x^{(i)}$ denotes the i-th component of vector x. The fitting problem then becomes an optimization problem

$$\min_{\theta \in \mathbb{R}^D} \mathcal{L}(\theta).$$

As the number of parameters is typically very large (due to an overparameterized model), it is natural to expect that not all of them have the same impact on $\mathcal{L}$. Some of the parameters might greatly change the value of the loss function when modified while others might barely change its value, or even not at all. It is even possible that some combinations of the parameters does not change the loss function's value (for example, increasing a parameter's value while decreasing another one's value might have little or no effect at all on $\mathcal{L}$). We illustrate this with a toy example.

Example 3.4.1 Imagine we are given a dataset

$$\mathcal{D} = \{(x_i, y_i\}_{i=1}^n$$

with $x_i \in \mathbb{R}^2$ and $y_i \in \mathbb{R}$. We want to fit a model

$$f_\theta(x) = \theta^{(1)} x^{(1)} + \theta^{(2)} x^{(2)}$$

for parameters $\theta = (\theta^{(1)}, \theta^{(2)})^\top \in \mathbb{R}^2$. Additionally, suppose that the features $x_i(1)$ and $x_i^{(2)}$ are actually collinear in such a way that $x_i^{(2)} = x_i^{(1)} + \varepsilon$ where $\varepsilon > 0$ is very small. The training set loss using the MSE then becomes

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} ((\theta^{(1)} + \theta^{(2)}) x_i^{(1)} + \varepsilon \theta^{(2)} - y_i)^2. \tag{3.18}$$

We can now notice an anisotropic structure in $\mathcal{L}$. In fact, looking at (3.18), we can see that $\mathcal{L}$ depends mainly on $\theta^{(1)} + \theta^{(2)}$ rather than $\theta^{(1)}$ and $\theta^{(2)}$ separately, due to the small value of ε.

3.4.2 Anisotropic structure in neural networks

We are given again a dataset

$$\mathcal{D} := \{(x_i, y_i)\}_{i=1}^n$$

with n features $x_i \in \mathbb{R}^d$ and labels $y_i \in \mathbb{R}^K$ represented with one-hot encoding (this means that $y_i = e_j$ if the label y_i corresponds to class $j \in \{1, \ldots, K\}$). We consider

neural networks with d input features, one hidden layer composed of k hidden neurons and K outputs as illustrated in Figure 3.20.

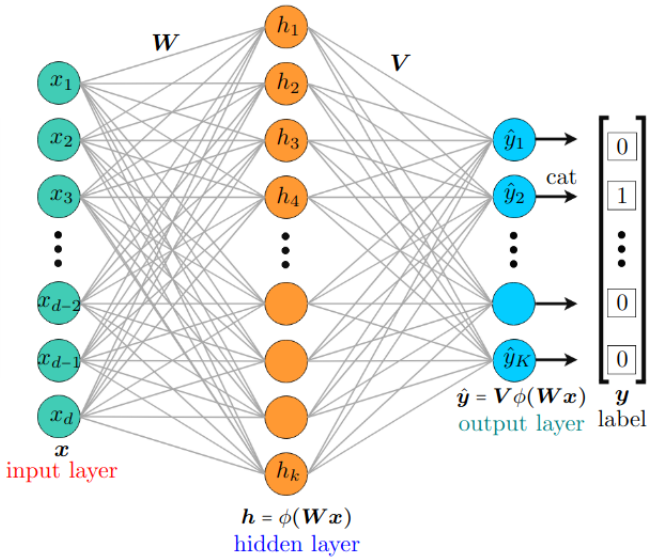


Figure 3.20: (Taken from [24]) Illustration of the architecture of the neural network defined in this section

Denote by $W \in \mathbb{R}^{k\times d}$ and $V \in \mathbb{R}^{K\times k}$ respectively the input-to-hidden and hidden-to-output weights. The relation between the input and the output of this neural network for weights W is a function

$$f(\cdot; W) : \mathbb{R}^d \to \mathbb{R}^K$$

defined by

$$f(x; W) := V\phi(Wx)$$

with $\phi : \mathbb{R}^k \to \mathbb{R}^k$ an activation function (e.g. softplus activation).

Remark For the rest of this section, we consider the hidden-to-input weights V to be fixed for clarity of exposition, and optimize over W. As stated in [24], this assumption does not reduce the generality of the following results.

We want the neural network to learn the dataset $\mathcal{D}$. In our setting, this amounts to solving the following optimization problem:

$$\min_{W\in\mathbb{R}^{k\times d}} \mathcal{L}(W) := \frac{1}{2}\sum_{i=1}^{n} \|f(x_i; W) - y_i\|^2.$$

Concatenating the y_i's into $y \in \mathbb{R}^{nK}$, the $f(x_i; W)$'s into $f(W) \in \mathbb{R}^{nK}$ and aggregating the weights W into one vector $w \in \mathbb{R}^{kd}$, the problem reduces to

$$\min_{w \in \mathbb{R}^{kd}} \mathcal{L}(w) := \frac{1}{2}\|f(w) - y\|^2.$$

One can compute the gradient of $\mathcal{L}$ with respect to w, which is given by

$$\nabla \mathcal{L}(w) = \mathcal{J}^\top(w) r(w) \tag{3.19}$$

where

$$\mathcal{J}(w) := \frac{\partial f(w)}{\partial w}$$

is the Jacobian of f with respect to w and

$$r(w) := f(w) - y$$

is called the residual vector (or misfit). We immediately see from (3.19) that the behavior of $\mathcal{L}$ depends on the spectrum of $\mathcal{J}$ as well as the residual vector r. As demonstrated numerically in [24], the Jacobian $\mathcal{J}$ associated with neural networks typically exhibits low-rank structure with a small number of large singular values, the rest of the singular values having small values. If we denote by $\mathcal{T}$ the linear subspace spanned by the singular vectors associated with the large singular values, then $\mathcal{T}^\top$ is the subspace spanned by the singular vectors associated with the smaller singular values and as the gradient of a function gives us information on the rate of change of a function along a certain direction, we expect $\mathcal{L}$ to vary slowly along $\mathcal{T}^\top$. This suggests that $\mathcal{L}$ is a function with low anisotropic dimensionality.

Chapter 4

Conclusion

In this work, we investigated the use of random embeddings for high-dimensional optimization problems with special objectives, namely functions with anisotropic dimensionality, or AD functions. These functions generalize the notion of functions with low effective dimensionality that were already studied in [10]. Following the principal lines of their work, we derived lower bounds on the probability that the randomized reduced problem (RP$\mathcal{X}$) yields an ε-minimizer of (P). By replacing the isotropic model of approximate minimizers (a ball) with an anisotropic model (an ellipsoid), we were able to more accurately capture the directional variations inherent to AD functions. The obtained lower bounds were expressed in terms of three geometric quantities: conic intrinsic volumes, statistical dimension and Gaussian width. While no tractable way of computing the conic intrinsic volumes was found, the latter two quantities offered practical bounds. Indeed, by expressing the elliptical cone tangent to the ellipsoid as a linear transformation of a simpler circular cone, we were able to efficiently compute these quantities in order to derive general results from these bounds. We also analyzed the impact of the various parameters present in (RP$\mathcal{X}$), which aligned with our expectations. Since functions with low effective dimensionality are special AD functions with $L_2 = 0$, our results naturally recovered some already known behavior about these functions, thereby validating our analysis. Numerical evidence further suggested a generalization of Theorem 3.3.1, namely Conjecture 3.3.1, as the statistical dimension $\delta(\mathcal{C}_{\mathcal{E}_E(\Delta)})$ approaches $D - d_a$ well before L_2 vanishes. Finally, we observed that AD functions are not merely theoretical constructs but naturally arise in real-world optimization problems, including machine learning and deep learning. Our results offered more insight into the empirical success of random embeddings in high-dimensional problems, especially when the objective exhibits special structure. We believe these findings could provide a foundation for designing more efficient algorithms tailored to exploit such structural properties.

Bibliography

[1] Semyon Alesker, Joseph HG Fu, Eduardo Gallego, and Gil Solanes. *Integral geometry and valuations*. Springer, 2014.

[2] Dennis Amelunxen. *Geometric analysis of the condition of the convex feasibility problem*. PhD thesis, Universitätsbibliothek, 2011.

[3] Dennis Amelunxen. Measures on polyhedral cones: characterizations and kinematic formulas. *arXiv preprint arXiv:1412.1569*, 2014.

[4] Dennis Amelunxen and Martin Lotz. Gordon's inequality and condition numbers in conic optimization. *arXiv preprint arXiv:1408.3016*, 2014.

[5] Dennis Amelunxen and Martin Lotz. Intrinsic volumes of polyhedral cones: a combinatorial perspective. *Discrete & Computational Geometry*, 58:371–409, 2017.

[6] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

[7] Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[8] Coralia Cartis, Xinzhu Liang, Estelle Massart, and Adilet Otemissov. Learning the subspace of variation for global optimization of functions with low effective dimension. *arXiv preprint arXiv:2401.17825*, 2024.

[9] Coralia Cartis, Estelle Massart, and Adilet Otemissov. Constrained global optimization of functions with low effective dimensionality using multiple random embeddings. *arXiv preprint arXiv:2009.10446*, 2020.

[10] Coralia Cartis, Estelle Massart, and Adilet Otemissov. Global optimization using random embeddings. *Mathematical Programming*, 200(2):781–829, 2023.

[11] Coralia Cartis and Adilet Otemissov. A dimensionality reduction technique for unconstrained global optimization of functions with low effective dimensionality. *Information and Inference: A Journal of the IMA*, 11(1):167–201, 2022.

[12] Paul G Constantine. *Active subspaces: Emerging ideas for dimension reduction in parameter studies*. SIAM, 2015.

[13] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

[14] Yehoram Gordon. On milman's inequality and random subspaces which escape through a mesh in r n. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 1986–87*, pages 84–106. Springer, 1988.

[15] Martin Henk and María A Hernández Cifre. Intrinsic volumes and successive radii. *Journal of mathematical analysis and applications*, 343(2):733–742, 2008.

[16] Daniel Hug, Wolfgang Weil, et al. *Lectures on convex geometry*, volume 286. Springer, 2020.

[17] Zakhar Kabluchko and Dmitry Zaporozhets. Intrinsic volumes of sobolev balls with applications to brownian convex hulls. *Transactions of the American Mathematical Society*, 368(12):8873–8899, 2016.

[18] Yue Lu and Jein-Shan Chen. The variational geometry, projection expression and decomposition associated with ellipsoidal cones. *Journal of Nonlinear and Convex Analysis*, 20(4):715–738, 2019.

[19] Dan Margalit, Joseph Rabinoff, and Larry Rolen. Interactive linear algebra. *Georgia Institute of Technology*, page 18, 2017.

[20] Michael B McCoy and Joel A Tropp. From steiner formulas for cones to concentration of intrinsic volumes. *Discrete & Computational Geometry*, 51:926–963, 2014.

[21] Michael Brian McCoy. *A geometric analysis of convex demixing*. California Institute of Technology, 2013.

[22] Sergii Myroshnychenko, Kateryna Tatarko, and Vladyslav Yaskin. Unique determination of ellipsoids by their dual volumes. *International Mathematics Research Notices*, 2022(17):13569–13589, 2022.

[23] Adilet Otemissov. *Dimensionality reduction techniques for global optimization.* PhD thesis, University of Oxford, 2020.

[24] Samet Oymak, Zalan Fabian, Mingchen Li, and Mahdi Soltanolkotabi. Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian. *arXiv preprint arXiv:1906.05392*, 2019.

[25] Grigoris Paouris and Peter Pivovarov. Intrinsic volumes and linear contractions. *Proceedings of the American Mathematical Society*, pages 1805–1808, 2013.

[26] Luis A Santaló. *Integral geometry and geometric probability.* Cambridge university press, 2004.

[27] Rolf Schneider. *Convex bodies: the Brunn–Minkowski theory*, volume 151. Cambridge university press, 2013.

[28] Rolf Schneider and Wolfgang Weil. *Stochastic and integral geometry*, volume 1. Springer, 2008.

[29] Zhen Shao. *On random embeddings and their application to optimisation.* University of Oxford (United Kingdom), 2021.

[30] Roman Vershynin. Estimation in high dimensions: a geometric perspective. In *Sampling Theory, a Renaissance: Compressive Sensing and Other Developments*, pages 3–66. Springer, 2015.

[31] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[32] RA Vitale. On the gaussian representation of intrinsic volumes. *Statistics & probability letters*, 78(10):1246–1249, 2008.

[33] Ziyu Wang, Frank Hutter, Masrour Zoghi, David Matheson, and Nando De Feitas. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55:361–387, 2016.

[34] Yuting Wei, Martin J Wainwright, and Adityanand Guntuboyina. The geometry of hypothesis testing over convex cones: Generalized likelihood ratio tests and minimax radii. 2019.

[35] JC Zhou and Jein-Shan Chen. Properties of circular cone and spectral factorization associated with circular cone. *J. Nonlinear Convex Anal*, 14(4):807–816, 2013.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**