Approximate Nearest Neighbor Searching with Non-Euclidean and Weighted Distances

Ahmed Abdelkader*
Department of Computer Science
University of Maryland, College Park, Maryland 20742
akader@cs.umd.edu

Sunil Arya†
Department of Computer Science and Engineering
The Hong Kong University of Science and Technology, Hong Kong
arya@cse.ust.hk

Guilherme D. da Fonseca‡
Université Clermont Auvergne, LIMOS, and INRIA Sophia Antipolis, France
fonseca@isima.fr

David M. Mount*
Department of Computer Science and Institute for Advanced Computer Studies
University of Maryland, College Park, Maryland 20742
mount@umd.edu

Abstract

We present a new approach to ε-approximate nearest-neighbor queries in fixed dimension under a variety of non-Euclidean distances. We consider two families of distance functions: (a) convex scaling distance functions including the Mahalanobis distance, the Minkowski metric and multiplicative weights, and (b) Bregman divergences including the Kullback-Leibler divergence and the Itakura-Saito distance.

As the fastest known data structures rely on the *lifting transformation*, their application is limited to the Euclidean metric, and alternative approaches for other distance functions are much less efficient. We circumvent the reliance on the lifting transformation by a careful application of *convexification*, which appears to be relatively new to computational geometry.

We are given n points in $\mathbb{R}^d$, each a site possibly defining its own distance function. Under mild assumptions on the growth rates of these functions, the proposed data structures answer queries in logarithmic time using $O(n \log(1/\varepsilon)/\varepsilon^{d/2})$ space, which nearly matches the best known results for the Euclidean metric.

1 Introduction

Nearest-neighbor searching is a fundamental retrieval problem with numerous applications in fields such as machine learning, data mining, data compression, and pattern recognition. A set of n points,

*Research supported by NSF grant CCF–1618866.

†Research supported by the Research Grants Council of Hong Kong, China under project number 16200014.

‡Research supported by the European Research Council under ERC Grant Agreement number 339025 GUDHI (Algorithmic Foundations of Geometric Understanding in Higher Dimensions).

called *sites*, is preprocessed into a data structure such that, given any query point q, it is possible to report the site that is closest to q. The most common formulation involves points in $\mathbb{R}^d$ under the Euclidean metric. Unfortunately, the best solution achieving $O(\log n)$ query time uses roughly $O(n^{d/2})$ storage space [20], which is too high for many applications.

This has motivated the study of approximations. Given an approximation parameter $\varepsilon > 0$, *ε-approximate nearest-neighbor searching* (ε-ANN) returns any site whose distance from q is within a factor of $1+\varepsilon$ of the distance to the true nearest neighbor. Throughout, we focus on $\mathbb{R}^d$ for fixed d and on data structures that achieve logarithmic query time of $O(\log \frac{n}{\varepsilon})$. The objective is to produce data structures of linear storage while minimizing the dependencies on ε, which typically grow exponentially with the dimension. Har-Peled showed that logarithmic query time could be achieved for Euclidean ε-ANN queries using roughly $O(n/\varepsilon^d)$ space through the *approximate Voronoi diagram* (AVD) data structure [24]. Despite subsequent work on the problem (see, e.g., [8, 10]), the storage requirements needed to achieve logarithmic query time remained essentially unchanged for over 15 years.

Recently, the authors [5,6] succeeded in reducing the storage to $O(n/\varepsilon^{d/2})$ by applying techniques from convex approximation.[1] Unlike the simpler data structure of [24], which can be applied to a variety of metrics, this recent data structure exploits properties that are specific to Euclidean space, which significantly limits its applicability. In particular, it applies a reduction to approximate polytope membership [8] based on the well-known *lifting transformation* [22]. However, this transformation applies only for the Euclidean distance.

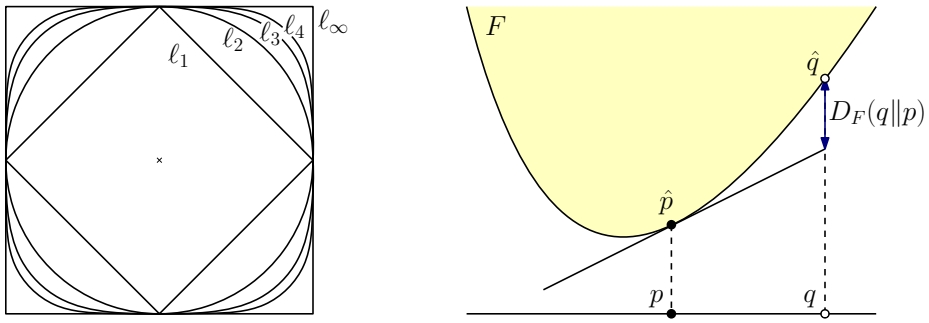
Note that all aforementioned data structures rely on the triangle inequality. Therefore, they cannot generally be applied to situations where each site is associated with its own distance function as arises, for example, with multiplicatively weighted sites (defined below).

Har-Peled and Kumar introduced a powerful technique to overcome this limitation through the use of *minimization diagrams* [25]. For each site p_i, let $f_i : \mathbb{R}^d \to \mathbb{R}^+$ be the associated distance function. Let $\mathcal{F}_{\min}$ denote the pointwise minimum of these functions, that is, the *lower-envelope function*. Clearly, approximating the value of $\mathcal{F}_{\min}$ at a query point q is equivalent to approximating the distance to q's nearest neighbor.[2] Har-Peled and Kumar proved that ε-ANN searching over a wide variety of distance functions (including additively and multiplicatively weighted sites) could be cast in this manner [25]. They formulated this problem in a very abstract setting, where no explicit reference is made to sites. Instead the input is expressed in terms of abstract properties of the distance functions, such as their growth rates and "sketchability." While this technique is very general, the complexity bounds are much worse than for the corresponding concrete versions. For example, in the case of Euclidean distance with multiplicative weights, in order to achieve logarithmic query time, the storage used is $O((n \log^{d+2} n)/\varepsilon^{2d+2} + n/\varepsilon^{d^2+d})$. Similar results are achieved for a number of other distance functions that are considered in [25].

This motivates the question of whether it is possible to answer ANN queries for non-Euclidean distance functions while matching the best bounds for Euclidean ANN queries. In this paper, we present a general approach for designing such data structures achieving $O(\log \frac{n}{\varepsilon})$ query time and $O((n/\varepsilon^{d/2}) \log \frac{1}{\varepsilon})$ storage. Thus, we suffer only an extra $\log \frac{1}{\varepsilon}$ factor in the space compared to the best results for Euclidean ε-ANN searching. We demonstrate the power of our approach by applying it to a number of natural problems:

[1] Chan [18] presented a similar result by a very different approach, and it generalizes to some other distance functions, however the query time is not logarithmic.

[2] The idea of using envelopes of functions for the purpose of nearest-neighbor searching has a long history, and it is central to the well-known relationship between the Euclidean Voronoi diagram of a set of points in $\mathbb{R}^d$ and the lower envelope of a collection of hyperplanes in $\mathbb{R}^{d+1}$ through the lifting transformation [22].



(a) (b)

Figure 1: (a) Unit balls in different Minkowski norms. (b) Geometric interpretation of the Bregman divergence.

Minkowski Distance: The ℓ_k distance (see Figure 1(a)) between two points p and q is defined as $\|q-p\|_k = (\sum_{i=1}^{d} |p_i - q_i|^k)^{\frac{1}{k}}$. Our results apply for any real constant $k > 1$.

Multiplicative Weights: Each site p is associated with weight $w_p > 0$ and $f_p(q) = w_p\|q-p\|$. The generalization of the Voronoi diagram to this distance function is known as the *Möbius diagram* [15]. Our results generalize from ℓ_2 to any Minkowski ℓ_k distance, for constant $k > 1$.

Mahalanobis Distance: Each site p is associated with a $d \times d$ positive-definite matrix M_p and $f_p(q) = \sqrt{(p-q)^\intercal M_p (p-q)}$. Mahalanobis distances are widely used in machine learning and statistics. Our results hold under the assumption that for each point p, the ratio between the maximum and minimum eigenvalues of M_p is bounded.

Scaling Distance Functions: Each site p is associated with a closed convex body K_p whose interior contains the origin, and $f_p(q)$ is the smallest r such that $(q-p)/r \in K_p$ (or zero if $q = p$). (These are also known as *convex distance functions* [19].) These generalize and customize normed metric spaces by allowing metric balls that are not centrally symmetric and allowing each site to have its own distance function.

Scaling distance functions generalize the Minkowski distance, multiplicative weights, and the Mahalanobis distance. Our results hold under the assumption that the convex body K_p inducing the distance function satisfies certain assumptions. First, it needs to be *fat* in the sense that it can be sandwiched between two Euclidean balls centered at the origin whose radii differ by a constant factor. Second, it needs to be *smooth* in the sense that the radius of curvature for every point on K_p's boundary is within a constant factor of its diameter. (Formal definitions will be given in Section 4.2.)

Theorem 1.1 (ANN for Scaling Distances)**.** *Given an approximation parameter $0 < \varepsilon \leq 1$ and a set S of n sites in $\mathbb{R}^d$ where each site $p \in S$ is associated with a fat, smooth convex body $K_p \subset \mathbb{R}^d$ (as defined above), there exists a data structure that can answer ε-approximate nearest-neighbor queries with respect to the respective scaling distance functions defined by K_p with*

$$\textit{Query time: } O\left(\log \frac{n}{\varepsilon}\right) \quad \textit{and} \quad \textit{Space: } O\left(\frac{n \log \frac{1}{\varepsilon}}{\varepsilon^{d/2}}\right).$$

Another important application that we consider is the Bregman divergence. Bregman divergences generalize the squared Euclidean distance [16], the Kullback-Leibler divergence (also known as relative entropy) [27], and the Itakura-Saito distance [26] among others. They have numerous applications in machine learning and computer vision [13, 30].

Bregman Divergence: Given an open convex domain $\mathcal{X} \subseteq \mathbb{R}^d$, a strictly convex and differentiable real-valued function F on $\mathcal{X}$, and $q, p \in \mathcal{X}$, the *Bregman divergence* of q from p is

$$D_F(q,p) \;=\; F(q) - (F(p) + \nabla F(p) \cdot (q-p)).$$

where ∇F denotes the gradient of F and "$\cdot$" is the standard dot product.

The Bregman divergence has the following geometric interpretation (see Figure 1(b)). Let $\hat{p}$ denote the vertical projection of p onto the graph of F, that is, $(p, F(p))$, and define $\hat{q}$ similarly. $D_F(q,p)$ is the vertical distance between $\hat{q}$ and the hyperplane tangent to F at the point $\hat{p}$. Equivalently, $D_F(q,p)$ is just the error that results by estimating $F(q)$ by a linear model at p.

The Bregman divergence possibly lacks many of the properties of typical distance functions. It is generally not symmetric, that is, $D_F(q,p) \neq D_F(p,q)$, and it generally does not satisfy the triangle inequality, but it is a convex function in its first argument. Throughout, we treat the first argument q as the query point and the second argument p as the site, but it is possible to reverse these through dualization [16].

Data structures have been presented for answering exact nearest-neighbor queries in the Bregman divergence by Cayton [17] and Nielson *et al.* [28], but no complexity analysis was given. Worst-case bounds have been achieved by imposing restrictions on the function F. Various different complexity measures have been proposed, including the following. Given a parameter $\mu \geq 1$, and letting $\|p-q\|$ denote the Euclidean distance between p and q:

- D_F is *μ-asymmetric* if for all $p, q \in \mathcal{X}$, $D_F(q,p) \leq \mu D_F(p,q)$.
- D_F is *μ-similar*[3] if for all $p, q \in \mathcal{X}$, $\|q-p\|^2 \leq D_F(q,p) \leq \mu \|q-p\|^2$.

Abdullah *et al.* [1] presented data structures for answering ε-ANN queries for decomposable[4] Bregman divergences in spaces of constant dimension under the assumption of bounded similarity. Later, Abdullah and Venkatasubramanian [2] established lower bounds on the complexity of Bregman ANN searching under the assumption of bounded asymmetry.

Our results for ANN searching in the Bregman divergence are stated below. They hold under a related measure of complexity, called *τ-admissibility*, which is more inclusive (that is, weaker) than μ-similarity, but seems to be more restrictive than μ-asymmetry. It is defined in Section 5.1, where we also explore the relationships between these measures.

Theorem 1.2 (ANN for Bregman Divergences)**.** *Given a τ-admissible Bregman divergence D_F for a constant τ defined over an open convex domain $\mathcal{X} \subseteq \mathbb{R}^d$, a set S of n sites in $\mathbb{R}^d$, and an approximation parameter $0 < \varepsilon \leq 1$, there exists a data structure that can answer ε-approximate nearest-neighbor queries with respect to D_F with*

$$\textit{Query time: } O\Big(\log \frac{n}{\varepsilon}\Big) \quad \textit{and} \quad \textit{Space: } O\left(\frac{n \log \frac{1}{\varepsilon}}{\varepsilon^{d/2}}\right).$$

[3]Our definition of μ-similarity differs from that of [3]. First, we have replaced $1/\mu$ with μ for compatibility with asymmetry. Second, their definition allows for any Mahalanobis distance, not just Euclidean. This is a trivial distinction in the context of nearest-neighbor searching, since it is possible to transform between such distances by applying an appropriate positive-definite linear transformation to the query space. Lemma 6.4 shows that the result is a Bregman divergence.

[4]The sum of one-dimensional Bregman divergences.

Note that our results are focused on the *existence* of these data structures, and construction is not discussed. While we see no significant impediments to their efficient construction by modifying the constructions of related data structures, a number of technical results would need to be developed. We therefore leave the question of efficient construction as a rather technical but nonetheless important open problem.

1.1 Methods

Our solutions are all based on the application of a technique, called *convexification*. Recently, the authors showed how to efficiently answer several approximation queries with respect to convex polytopes [5–7], including polytope membership, ray shooting, directional width, and polytope intersection. As mentioned above, the linearization technique using the lifting transformation can be used to produce convex polyhedra for the sake of answering ANN queries, but it is applicable only to the Euclidean distance (or more accurately the squared Euclidean distance and the related power distance [12]). In the context of approximation, polytopes are not required. The convex approximation methods described above can be adapted to work on any convex body, even one with curved boundaries. This provides us with an additional degree of flexibility. Rather than applying a transformation to linearize the various distance functions, we can go a bit overboard and "convexify" them.

Convexification techniques have been used in non-linear optimization for decades [14], for example the αBB optimization method locally convexifies constraint functions to produce constraints that are easier to process [4]. However, we are unaware of prior applications of this technique in computational geometry in the manner that we use it. (For an alternate use, see [21].)

The general idea involves the following two steps. First, we apply a quadtree-like approach to partition the query space (that is, $\mathbb{R}^d$) into cells so that the restriction of each distance function within each cell has certain "nice" properties, which make it possible to establish upper bounds on the gradients and the eigenvalues of their Hessians. We then add to each function a common "convexifying" function whose Hessian has sufficiently small (in fact negative) eigenvalues, so that all the functions become concave (see Figure 3 in Section 3 below). We then exploit the fact that the lower envelope of concave functions is concave. The region lying under this lower envelope can be approximated by standard techniques, such as the ray-shooting data structure of [6]. We show that if the distance functions satisfy our admissibility conditions, this can be achieved while preserving the approximation errors.

The rest of the paper is organized as follows. In the next section we present definitions and preliminary results. Section 3 discusses the concept of convexification, and how it is applied to vertical ray shooting on the minimization diagram of sufficiently well-behaved functions. In Section 4, we present our solution to ANN searching for scaling distance functions, proving Theorem 1.1. In Section 5, we do the same for the case of Bregman divergence, proving Theorem 1.2. Finally, in Section 6 we present technical details that have been deferred from the main body of the paper.

2 Preliminaries

In this section we present a number of definitions and results that will be useful throughout the paper.

2.1 Notation and Assumptions

For the sake of completeness, let us recall some standard definitions. Given a function $f : \mathbb{R}^d \to \mathbb{R}$, its *graph* is the set of $(d+1)$-dimensional points $(x, f(x))$, its *epigraph* is the set of points on or above the graph, and its *hypograph* is the set of points on or below the graph (where the $(d+1)$-st axis is directed upwards). The *level set* (also called *level surface* if $d \geq 3$) of f is the set of points $x \in \mathbb{R}^d$ for which f has the same value.

The gradient and Hessian of a function generalize the concepts of the first and second derivative to a multidimensional setting. The *gradient* of f, denoted ∇f, is defined as the vector field $(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_d})^\mathsf{T}$. The gradient vector points in a direction in which the function grows most rapidly, and it is orthogonal to the level surface. For any point x and any unit vector v, the rate of change of f along v is given by the dot product $\nabla f(x) \cdot v$. The *Hessian* of f at x, denoted $\nabla^2 f(x)$, is a $d \times d$ matrix of second-order partial derivatives at x. For twice continuously differentiable functions, $\nabla^2 f(x)$ is symmetric, implying that it has d (not necessarily distinct) real eigenvalues.

Given a d-vector v, let $\|v\|$ denote its length under the *Euclidean norm*, and the *Euclidean distance* between points p and q is $\|q - p\|$. Given a $d \times d$ matrix A, its *spectral norm* is $\|A\| = \sup\{\|Ax\| / \|x\| \ : \ x \in \mathbb{R}^d \text{ and } x \neq 0\}$. Since the Hessian is a symmetric matrix, it follows that $\|\nabla^2 f(x)\|$ is the largest absolute value attained by the eigenvalues of $\nabla^2 f(x)$.

A real-valued function f defined on a nonempty subset $\mathcal{X}$ of $\mathbb{R}^d$ is *convex* if the domain $\mathcal{X}$ is convex and for any $x, y \in \mathcal{X}$ and $\alpha \in [0, 1]$, $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y)$, and it is *concave* if $-f$ is convex. A twice continuously differentiable function on a convex domain is convex if and only if its Hessian matrix is positive semidefinite in the interior of the domain. It follows that all the eigenvalues of the Hessian of a convex function are nonnegative.

Given a function $f : \mathbb{R}^d \to \mathbb{R}$ and a closed Euclidean ball B (or generally any closed bounded region), let $f^+(B)$ and $f^-(B)$ denote the maximum and minimum values, respectively, attained by $f(x)$ for $x \in B$. Similarly, define $\|\nabla f^+(B)\|$ and $\|\nabla^2 f^+(B)\|$ to be the maximum values of the norms of the gradient and Hessian, respectively, for any point in B.

2.2 Minimization Diagrams and Vertical Ray Shooting

Consider a convex domain $\mathcal{X} \subseteq \mathbb{R}^d$ and a set of functions $\mathcal{F} = \{f_1, \ldots, f_m\}$, where $f_i : \mathcal{X} \to \mathbb{R}^+$. Let $\mathcal{F}_{\min}$ denote the associated *lower-envelope function*, that is $\mathcal{F}_{\min}(x) = \min_{1 \leq i \leq m} f_i(x)$. As Har-Peled and Kumar [25] observed, for any $\varepsilon > 0$, we can answer ε-ANN queries on any set S by letting f_i denote the distance function to the ith site, and computing any index i (called a *witness*) such that $f_i(q) \leq (1+\varepsilon)\mathcal{F}_{\min}(q)$.

We can pose this as a geometric approximation problem in one higher dimension. Consider the hypograph in $\mathbb{R}^{d+1}$ of $\mathcal{F}_{\min}$, and let us think of the $(d+1)$st axis as indicating the *vertical* direction. Answering ε-ANN queries in the above sense can be thought of as approximating the result of a vertical ray shot upwards from the point $(q, 0) \in \mathbb{R}^{d+1}$ until it hits the lower envelope, where the allowed approximation error is $\varepsilon \mathcal{F}_{\min}(q)$. Because the error is relative to the value of $\mathcal{F}_{\min}(q)$, this is called a *relative ε-AVR query*. It is also useful to consider a variant in which the error is absolute. An *absolute ε-AVR query* returns any witness i such that $f_i(q) \leq \varepsilon + \mathcal{F}_{\min}(q)$ (see Fig. 2).

The hypograph of a general minimization diagram can be unwieldy. Our approach to answer AVR queries efficiently will involve subdividing space into regions such that within each region it is possible to transform the hypograph into a convex shape. In the next section, we will describe this transformation. Given this, our principal utility for answering ε-AVR queries efficiently is encapsulated in the following lemma (see Figure 2). The proof has been deferred to Section 6.

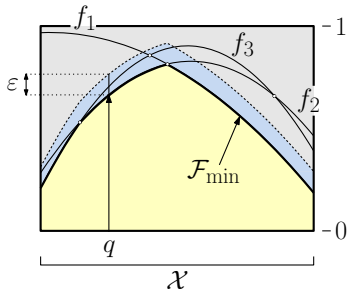


Figure 2: Approximate AVR query assuming absolute errors. For the query q, the exact answer is f_2, but f_3 would be acceptable.

Lemma 2.1. (Answering ε-AVR Queries) *Consider a unit ball $B \subseteq \mathbb{R}^d$ and a family of concave functions $\mathcal{F} = \{f_1, \ldots, f_m\}$ defined over B such that for all $1 \leq i \leq m$ and $x \in B$, $f_i(x) \in [0,1]$ and $\|\nabla f_i(x)\| \leq 1$. Then, for any $0 < \varepsilon \leq 1$, there is a data structure that can answer absolute ε-AVR queries in time $O(\log \frac{1}{\varepsilon})$ and storage $O((\frac{1}{\varepsilon})^{d/2})$.*

3 Convexification

In this section we discuss the key technique underlying many of our results. As mentioned above, our objective is to answer ε-AVR queries with respect to the minimization diagram, but this is complicated by the fact that it does not bound a convex set.

In order to overcome this issue, let us make two assumptions. First, we restrict the functions to a bounded convex domain, which for our purposes may be taken to be a closed Euclidean ball B in $\mathbb{R}^d$. Second, let us assume that the functions are smooth, implying in particular that each function f_i has a well defined gradient ∇f_i and Hessian $\nabla^2 f_i$ for every point of B. As mentioned above a function f_i is convex (resp., concave) over B if and only if all the eigenvalues of $\nabla^2 f_i(x)$ are nonnegative (resp., nonpositive). Intuitively, if the functions f_i are sufficiently well-behaved it is possible to compute upper bounds on the norms of the gradients and Hessians throughout B. Given $\mathcal{F}$ and B, let Λ^+ denote an upper bound on the largest eigenvalue of $\nabla^2 f_i(x)$ for any function $f_i \in \mathcal{F}$ and for any point $x \in B$.

We will apply a technique called *convexification* from the field of nonconvex optimization [4,14]. If we add to f_i any function whose Hessian has a maximum eigenvalue at most $-\Lambda^+$, we will effectively "overpower" all the upward curving terms, resulting in a function having only nonpositive eigenvalues, that is, a concave function.[5] The lower envelope of concave functions is concave, and so techniques for convex approximation (such as Lemma 2.1) can be applied to the hypograph of the resulting lower-envelope function.

To make this more formal, let $p \in \mathbb{R}^d$ and $r \in \mathbb{R}$ denote the center point and radius of B, respectively. Define a function ϕ (which depends on B and Λ^+) to be

$$\phi(x) \;=\; \frac{\Lambda^+}{2}\left(r^2 - \sum_{j=1}^{d}(x_j - p_j)^2\right) \;=\; \frac{\Lambda^+}{2}(r^2 - \|x-p\|^2).$$

It is easy to verify that ϕ evaluates to zero along B's boundary and is positive within B's interior. Also, for any $x \in \mathbb{R}^d$, the Hessian of $\|x-p\|^2$ (as a function of x) is a $d \times d$ diagonal matrix $2I$, and

[5] While this intuition is best understood for convex functions, it can be applied whenever there is an upper bound on the maximum eigenvalue.

therefore $\nabla^2\phi(x) = -\Lambda^+ I$. Now, define

$$\begin{aligned}\widehat{f}_i(x) &= f_i(x) + \phi(x), \quad \text{for } 1 \le i \le m, \text{ and} \\ \widehat{F}_{\min}(x) &= \min_{1\le i\le m} \widehat{f}_i(x) \;=\; \mathcal{F}_{\min}(x) + \phi(x).\end{aligned}$$

Because all the functions are subject to the same offset at each point x, $\widehat{F}_{\min}$ preserves the relevant combinatorial structure of $\mathcal{F}_{\min}$, and in particular f_i yields the minimum value to $\mathcal{F}_{\min}(x)$ at some point x if and only if $\widehat{f}_i$ yields the minimum value to $\widehat{F}_{\min}(x)$. Absolute vertical errors are preserved as well. Observe that $\widehat{F}_{\min}(x)$ matches the value of $\mathcal{F}_{\min}$ along B's boundary and is larger within its interior. Also, since $\nabla^2\phi(x) = -\Lambda^+ I$, it follows from elementary linear algebra that each eigenvalue of $\nabla^2\widehat{f}_i(x)$ is smaller than the corresponding eigenvalue of $\nabla^2 f_i(x)$ by Λ^+. Thus, all the eigenvalues of $\widehat{f}_i(x)$ are nonpositive, and so $\widehat{f}_i$ is concave over B. In turn, this implies that $\widehat{F}_{\min}$ is concave, as desired. We will show that, when properly applied, relative errors are nearly preserved, and hence approximating the convexified lower envelope yields an approximation to the original lower envelope.

3.1 A Short Example

As a simple application of this technique, consider the following problem. Let $\mathcal{F} = \{f_1, \ldots, f_m\}$ be a collection of m multivariate polynomial functions over $\mathbb{R}^d$ each of constant degree and having coefficients whose absolute values are $O(1)$ (see Figure 3(a)). It is known that the worst-case combinatorial complexity of the lower envelope of algebraic functions of fixed degree in $\mathbb{R}^d$ lies between $\Omega(n^d)$ and $O(n^{d+\alpha})$ for any $\alpha > 0$ [29], which suggests that any exact solution to computing a point on the lower envelope $\mathcal{F}_{\min}$ will either involve high space or high query time.

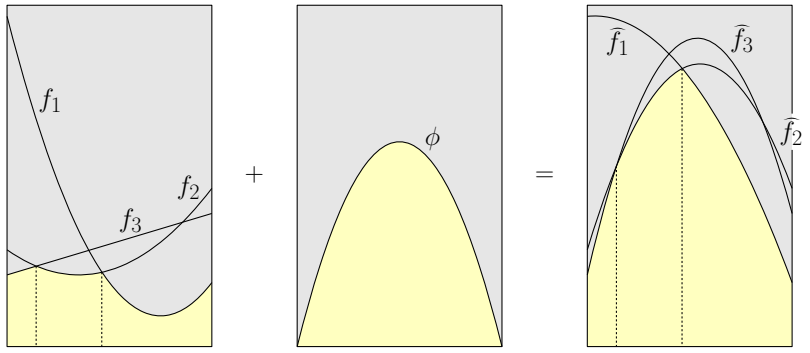


Figure 3: Convexification.

Let us consider a simple approximate formulation by restricting $\mathcal{F}$ to a unit d-dimensional Euclidean ball B centered at the origin. Given a parameter $\varepsilon > 0$, the objective is to compute for any query point $q \in \mathbb{R}^d$ an *absolute ε-approximation* by returning the index of a function f_i such that $f_i(q) \le \mathcal{F}_{\min}(q) + \varepsilon$. (While relative errors are usually desired, this simpler formulation is sufficient to illustrate how convexification works.) Since the degrees and coefficients are bounded, it follows that for each $x \in B$, the norms of the gradients and Hessians for each function f_i are bounded. A simple naive solution would be to overlay B with a grid with cells of diameter $\Theta(\varepsilon)$ and compute the answer for a query point centered within each grid cell. Because the gradients are

bounded, the answer to the query for the center point is an absolute ε-approximation for any point in the cell. This produces a data structure with space $O((\frac{1}{\varepsilon})^d)$.

To produce a more space-efficient solution, we apply convexification. Because the eigenvalues of the Hessians are bounded for all $x \in B$ and all functions f_i, it follows that there exists an upper bound $\Lambda^+ = O(1)$ on all the Hessian eigenvalues. Therefore, by computing the convexifying function ϕ described above (see Figure 3(b)) to produce the new function $\widehat{F}_{\min}$ (see Figure 3(c)) we obtain a concave function. It is easy to see that ϕ has bounded gradients and therefore so does $\widehat{F}_{\min}$. The hypograph of the resulting function when suitably trimmed is a convex body of constant diameter residing in $\mathbb{R}^{d+1}$. After a suitable scaling (which will be described later in Lemma 3.2), the functions can be transformed so that we may apply Lemma 2.1 to answer approximate vertical ray-shooting queries in time $O(\log \frac{1}{\varepsilon})$ with storage $O((\frac{1}{\varepsilon})^{d/2})$. This *halves* the exponential dependence in the dimension over the simple approach.

3.2 Admissible Distance Functions

A key question for us is whether the convexification process preserves approximation errors. We will show that if the functions satisfy certain admissibility properties, then this will be the case. We are given a domain $\mathcal{X} \subseteq \mathbb{R}^d$, and we assume that each distance function is associated with a defining site $p \in \mathcal{X}$. Consider a distance function $f_p : \mathcal{X} \to \mathbb{R}^+$ with a well-defined gradient and Hessian for each point of $\mathcal{X}$.[6] Given $\tau > 0$, we say that f_p is *τ-admissible* if for all $x \in \mathcal{X}$:

(*i*) $\|\nabla f_p(x)\| \, \|x - p\| \;\leq\; \tau f_p(x)$, and

(*ii*) $\|\nabla^2 f_p(x)\| \, \|x - p\|^2 \;\leq\; \tau^2 f_p(x)$.

Intuitively, an admissible function exhibits growth rates about the site that are polynomially upper bounded. For example, it is easy to prove that $f_p(x) = \|x - p\|^c$ is $O(c)$-admissible, for any $c \geq 1$.

Admissibility implies bounds on the magnitudes of the function values, gradients, and Hessians. Given a Euclidean ball B and site p, we say that B and p are *β-separated* if $\operatorname{dist}(p, B)/\operatorname{diam}(B) \geq \beta$ (where $\operatorname{dist}(p, B)$ is the minimum Euclidean distance between p and B and $\operatorname{diam}(B)$ B's diameter). The following lemma presents upper bounds on $f^+(B)$, $\|\nabla f^+(B)\|$, and $\|\nabla^2 f^+(B)\|$ in terms of these quantities. (Recall definitions from Section 2.1.) The proof is rather technical and has been deferred to Section 6.

Lemma 3.1. *Consider an open convex domain $\mathcal{X}$, a site $p \in \mathcal{X}$, a τ-admissible distance function f_p, and a Euclidean ball $B \subset \mathcal{X}$. If B and p are $(\tau\kappa)$-separated for $\kappa > 1$, then:*

(*i*) $f_p^+(B) \;\leq\; f_p^-(B)\kappa/(\kappa - 1)$,

(*ii*) $\|\nabla f_p^+(B)\| \;\leq\; f_p^+(B)/(\kappa \operatorname{diam}(B))$, *and*

(*iii*) $\|\nabla^2 f_p^+(B)\| \;\leq\; f_p^+(B)/(\kappa \operatorname{diam}(B))^2$.

For the special case of $\kappa = 2$, we obtain the following specific bounds.

Corollary 3.1. *Consider an open convex domain $\mathcal{X}$, a site $p \in \mathcal{X}$, a τ-admissible distance function f_p, and a Euclidean ball $B \subset \mathcal{X}$. If B and p are (2τ)-separated, then:*

(*i*) $f_p^+(B) \;\leq\; 2 f_p^-(B)$,

[6] This assumption is really too strong, since distance functions often have undefined gradients or Hessians at certain locations (e.g., the sites themselves). For our purposes it suffices that the gradient and Hessian are well defined at any point within the region where convexification will be applied.

(*ii*) $\|\nabla f_p^+(B)\| \leq f_p^+(B)/(2 \cdot \text{diam}(B))$, *and*

(*iii*) $\|\nabla^2 f_p^+(B)\| \leq f_p^+(B)/(2 \cdot \text{diam}(B))^2$.

3.3 Convexification and Ray Shooting

A set $\mathcal{F} = \{f_1, \ldots, f_m\}$ of τ-admissible functions is called a *τ-admissible family of functions*. Let $\mathcal{F}_{\min}$ denote the associated lower-envelope function. In Lemma 2.1 we showed that absolute ε-AVR queries could be answered efficiently in a very restricted context. This will need to be generalized the purposes of answering ANN queries, however.

The main result of this section states that if the sites defining the distance functions are sufficiently well separated from a Euclidean ball, then (through convexification) ε-AVR queries can be efficiently answered. The key idea is to map the ball and functions into the special structure required by Lemma 2.1, and to analyze how the mapping process affects the gradients and Hessians of the functions.

Lemma 3.2. (Convexification & Ray-Shooting) *Consider a Euclidean ball $B \in \mathbb{R}^d$ and a family of τ-admissible distance functions $\mathcal{F} = \{f_1, \ldots, f_m\}$ over B such that each associated site is (2τ)-separated from B. Given any $\varepsilon > 0$, there exists a data structure that can answer relative ε-AVR queries with respect to $\mathcal{F}_{\min}$ in time $O(\log \frac{1}{\varepsilon})$ with storage $O((\frac{1}{\varepsilon})^{d/2})$.*

Proof. We will answer approximate vertical ray-shooting queries by a reduction to the data structure given in Lemma 2.1 for answering approximate central ray-shooting queries. In order to apply this lemma, we need to transform the problem into the canonical form prescribed by that lemma.

We may assume without loss of generality that f_1 is the function that minimizes $f_1^-(B)$ among all the functions in $\mathcal{F}$. By Corollary 3.1(i), $f_1^+(B) \leq 2f_1^-(B)$. For all i, we may assume that $f_i^-(B) \leq 2f_1^-(B)$ for otherwise this function is greater than f_1 throughout B, and hence it does not contribute to $\mathcal{F}_{\min}$. Under this assumption, it follows that $f_i^+(B) \leq 4f_1^-(B)$.

In order to convert these functions into the desired form, define $h = 5f_1^-(B)$, $r = \text{radius}(B)$, and let $c \in \mathbb{R}^d$ denote the center of B. Let B_0 be a unit ball centered at the origin, and for any $x \in B_0$, let $x' = rx + c$. Observe that $x \in B_0$ if and only if $x' \in B$. For each i, define the normalized distance function

$$g_i(x) \;=\; \frac{f_i(x')}{h}.$$

We assert that these functions satisfy the following properties. They are straightforward consequences of admissibility and separation. For completeness, they are proved in Lemma 6.1 in Section 6.

(a) $g_i^+(B_0) \leq 4/5$ and $g_i^-(B_0) \geq 1/5$

(b) $\|\nabla g_i^+(B_0)\| \leq 1/2$

(c) $\|\nabla^2 g_i^+(B_0)\| \leq 1/4$

Next, we convexify these functions. To do this, define $\phi(x) = (1 - \|x\|^2)/8$. Observe that for any $x \in B_0$, $\phi(x) \in [0, 1/8]$ and $\|\nabla\phi(x)\| = \|x\|/4$ and $\nabla^2\phi(x)$ is the diagonal matrix $-(1/4)I$. Define

$$\widehat{g}_i(x) \;=\; g_i(x) + \phi(x).$$

It is easily verified that these functions satisfy the following properties.

(a) $\widehat{g}_i^+(B_0) \leq 1$ and $\widehat{g}_i^-(B_0) \geq 1/5$

(b) $\|\nabla \widehat{g}_i^+(B_0)\| \leq \|\nabla g_i^+(B_0)\| + \|\nabla \phi^+(B_0)\| < 1$

(c) $\|\nabla^2 \widehat{g}_i^+(B_0)\| \leq \|\nabla^2 g_i^+(B_0)\| - (1/4) \leq 0$

By property (c), these functions are concave over B_0. Given that $\widehat{g}_i^-(B_0) \geq 1/5$, in order to answer AVR queries to a relative error of ε, it suffices to answer AVR queries to an absolute error of $\varepsilon' = \varepsilon/5$. Therefore, we can apply Lemma 2.1 (using ε' in place of ε) to obtain a data structure that answers relative ε-AVR queries with respect to $\mathcal{F}_{\min}$ in time $O(\log \frac{1}{\varepsilon})$ with storage $O((\frac{1}{\varepsilon})^{d/2})$, as desired. □

Armed with this tool, we are now in a position to present the data structures for answering ε-ANN queries for each of our applications, which we do in the subsequent sections.

4 Answering ANN Queries for Scaling Distance Functions

Recall that in a scaling distance we are given a convex body K that contains the origin in its interior, and the distance from a query point q to a site p is defined to be zero if $p = q$ and otherwise it is the smallest r such that $(q - p)/r \in K$.[7] The body K plays the role of a unit ball in a normed metric, but we do not require that the body be centrally symmetric. In this section we establish Theorem 1.1 by demonstrating a data structure for answering ε-ANN queries given a set S of n sites, where each site p_i is associated with a scaling distance whose unit ball is a fat, smooth convex body.

Before presenting the data structure, we present two important preliminary results. The first, given in Section 4.1, explains how to subdivide space into a number of regions, called *cells*, that possess nice separation properties with respect to the sites. The second, given in Section 4.2, presents key technical properties of scaling functions whose unit balls are fat and smooth.

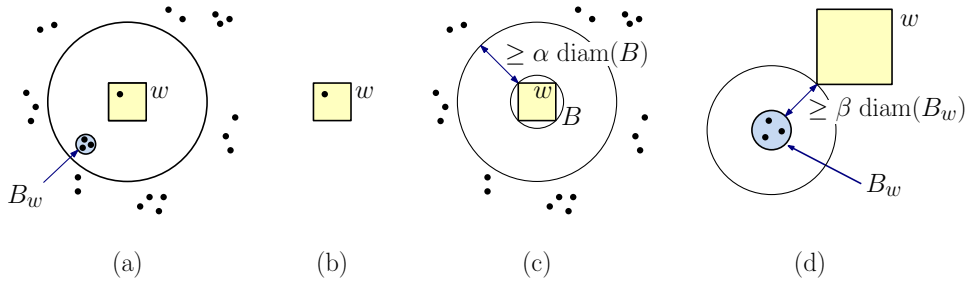


Figure 4: Basic separation properties for Lemma 4.1.

4.1 AVD and Separation Properties

In order to apply the convexification process, we will first subdivide space into regions, each of which satisfies certain separation properties with respect to the sites S. This subdivision results from a height-balanced variant of a quadtree, called a *balanced box decomposition tree* (or BBD tree) [11]. Each cell of this decomposition is either a quadtree box or the set-theoretic difference of two such boxes. Each leaf cell is associated with an auxiliary ANN data structure for the query points in the cell, and together the leaf cells subdivide all of $\mathbb{R}^d$.

[7] This can be readily generalized to squared distances, that is, the smallest r such that $(q - p)/\sqrt{r} \in K$. A relative error of $1 + \varepsilon$ in the squared distance, reduces to computing a $\sqrt{1 + \varepsilon}$ relative error in the original distance. Since $\sqrt{1 + \varepsilon} \approx (1 + \varepsilon/2)$ for small ε, our approach can be applied but with a slightly smaller value of ε. This generalizes to any constant power.

The separation properties are essentially the same as those of the AVD data structure of [10]. For any leaf cell w of the decomposition, the sites can be partitioned into three subsets, any of which may be empty (see Figure 4(a)). First, a single site may lie within w. Second, a subset of sites, called the *outer cluster*, is well-separated from the cell. Finally, there may be a dense cluster of points, called the *inner cluster*, that lie within a ball B_w that is well-separated from the cell. After locating the leaf cell containing the query point, the approximate nearest neighbor is computed independently for each of these subsets (by a method to be described later), and the overall closest is returned. The next lemma formalizes these separation properties. It follows easily from Lemma 6.1 in [9]. Given a BBD-tree cell w and a point $p \in \mathbb{R}^d$, let $\text{dist}(p, w)$ denote the minimum Euclidean distance from p to any point in w.

Lemma 4.1 (Basic Separation Properties)**.** *Given a set S of n points in $\mathbb{R}^d$ and real parameters $\alpha, \beta \geq 2$. It is possible to construct a BBD tree T with $O(\alpha^d n \log \beta)$ nodes, whose leaf cells cover $\mathbb{R}^d$ and for every site $p \in S$, either*

(*i*) *it lies within w, but there can be at most one site for which this holds (see Figure 4(b)),*

(*ii*) (outer cluster) *letting B denote the smallest Euclidean ball enclosing w,* $\text{dist}(p, B) \geq \alpha \cdot \text{diam}(B)$ *(see Figure 4(c)), or*

(*iii*) (inner cluster) *there exists a ball B_w associated with w such that* $\text{dist}(B_w, w) \geq \beta \cdot \text{diam}(B_w)$ *and $p \in B_w$ (see Figure 4(d)).*

Furthermore, it is possible to compute the tree T in total time $O(\alpha^d n \log n \log \beta)$, and the leaf cell containing a query point can be located in time $O(\log(\alpha n) + \log \log \beta)$

4.2 Admissibility for Scaling Distances

In this section we explore how properties of the unit ball affect the effectiveness of convexification. Recall from Section 3 that convexification relies on the admissibility of the distance function, and we show here that this will be guaranteed if unit balls are fat, well centered, and smooth.

Given a convex body K and a parameter $0 < \gamma \leq 1$, we say that K is *centrally γ-fat* if there exist Euclidean balls B and B' centered at the origin, such that $B \subseteq K \subseteq B'$, and $\text{radius}(B)/\text{radius}(B') \geq \gamma$. Given a parameter $0 < \sigma \leq 1$, we say that K is *σ-smooth* if for every point x on the boundary of K, there exists a closed Euclidean ball of diameter $\sigma \cdot \text{diam}(K)$ that lies within K and has x on its boundary. We say that a scaling distance function is a *(γ, σ)-distance* if its associated unit ball B is both centrally γ-fat and σ-smooth.

In order to employ convexification for scaling distances, it will be useful to show that smoothness and fatness imply that the associated distance functions are admissible. This is encapsulated in the following lemma. It follows from a straightforward but rather technical exercise in multivariate differential calculus. We include a complete proof in Section 6.

Lemma 4.2. *Given positive reals γ and σ, let f_p be a (γ, σ)-distance over $\mathbb{R}^d$ scaled about some point $p \in \mathbb{R}^d$. There exists τ (a function of γ and σ) such that f_p is τ-admissible.*

Our results on ε-ANN queries for scaling distances will be proved for any set of sites whose associated distance functions (which may be individual to each site) are all (γ, σ)-distances for fixed γ and σ. Our results on the Minkowski and Mahalanobis distances thus arise as direct consequences of the following easy observations.

Lemma 4.3.

(i) *For any positive real $k > 1$, the Minkowski distance ℓ_k is a (γ, σ)-distance, where γ and σ are functions of k and d.*

This applies to multiplicatively weighted Minkowski distances as well.

(ii) *The Mahalanobis distance defined by a matrix M_p is a (γ, σ)-distance, where γ and σ are functions of M_p's minimum and maximum eigenvalues.*

4.3 ANN Data Structure for Scaling Functions

Let us return to the discussion of how to answer ε-ANN queries for a family of (γ, σ)-distance functions. By Lemma 4.2, such functions are τ-admissible, where τ depends only on γ and σ.

We begin by building an (α, β)-AVD over $\mathbb{R}^d$ by invoking Lemma 4.1 for $\alpha = 2\tau$ and $\beta = 10\tau/\varepsilon$. (These choices will be justified below.) For each leaf cell w, the nearest neighbor of any query point $q \in w$ can arise from one of the three cases in the lemma. Case (i) is trivial since there is just one point.

Case (ii) (the *outer cluster*) can be solved easily by reduction to Lemma 3.2. Recall that we have a BBD-tree leaf cell w, and the objective is to compute an ε-ANN from among the outer cluster, that is, a set whose sites are at Euclidean distance at least $\alpha \cdot \text{diam}(w)$ from w. Let B denote the smallest Euclidean ball enclosing w and let $\mathcal{F}$ be the family of distance functions associated with the sites of the outer cluster. Since $\alpha = 2\tau$, B is (2τ)-separated from the points of the outer cluster. By Lemma 3.2, we can answer ε-AVR queries with respect to $\mathcal{F}_{\min}$, and this is equivalent to answering ε-ANN queries with respect to the outer cluster. The query time is $O(\log \frac{1}{\varepsilon})$ and the storage is $O((\frac{1}{\varepsilon})^{d/2})$.

All that remains is case (iii), the *inner cluster*. Recall that these sites lie within a ball B_w such that $\text{dist}(B_w, w) \geq \beta \cdot \text{diam}(B_w)$. In approximate Euclidean nearest-neighbor searching, a separation as large as β would allow us to replace all the points of B_w with a single representative site, but this is not applicable when different sites are associated with different scaling distance functions. We will show instead that queries can be answered by partitioning the query space into a small number of regions such that Lemma 3.2 can be applied to each region. Let $\{p_1, \ldots, p_m\}$ denote the sites lying within B_w, and let $\mathcal{F} = \{f_1, \ldots, f_m\}$ denote the associated family of (γ, σ)-distance functions.

Let p' be the center of B_w, and for $1 \leq i \leq m$, define the *perturbed distance function* $f'_i(x) = f_i(x + p_i - p')$ to be the function that results by moving p_i to p' without altering the unit metric ball. Let $\mathcal{F}'$ denote the associated family of distance functions. Our next lemma shows that this perturbation does not significantly alter the function values.

Lemma 4.4. *Let $p \in \mathbb{R}^d$ be the site of a τ-admissible distance function f. Let B be a ball containing p and let x be a point that is β-separated from B for $\beta \geq 2\tau$. Letting p' denote B's center, define $f'(x) = f(x + p - p')$. Then*

$$\frac{|f'(x) - f(x)|}{f(x)} \leq \frac{2\tau}{\beta}.$$

Proof. Define B_x to be the translate of B whose center coincides with x. Since p and p' both lie within B, x and $x+p-p'$ both lie within B_x. Let $\kappa = \beta/\tau$. Since x and B are β-separated, p' and B_x are also β-separated. Equivalently, they are $(\tau\kappa)$-separated. Because $\kappa \geq 2$, $\kappa/(\kappa - 1) \leq (1 + 2/\kappa)$. Because f' has the same unit metric ball as f, it is also τ-admissible, and so by Lemma 3.1

$$\begin{aligned} f'^{+}(B_x) &\leq \frac{\kappa}{\kappa - 1} f'^{-}(B_x) \leq \left(1 + \frac{2}{\kappa}\right) f'^{-}(B_x) \\ &= \left(1 + \frac{2\tau}{\beta}\right) f'^{-}(B_x). \end{aligned}$$

Letting $x' = x - (p - p')$, we have $f(x) = f'(x')$. Clearly $x' \in B_x$. Let us assume that $f'(x) \geq f(x)$. (The other case is similar.) We have

$$\begin{aligned} f'(x) - f(x) &= f'(x) - f'(x') \leq f'^{+}(B_x) - f'^{-}(B_x) \\ &\leq \frac{2\tau}{\beta} f'^{-}(B_x) \leq \frac{2\tau}{\beta} f'(x') = \frac{2\tau}{\beta} f(x), \end{aligned}$$

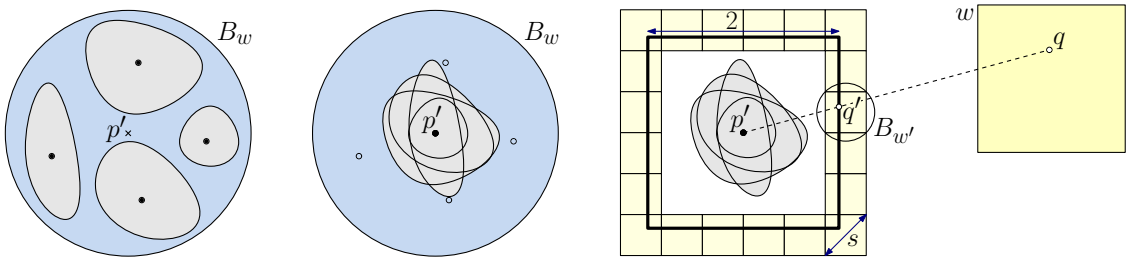
which implies the desired inequality. □

Since every point $x \in w$ is β-separated from B_w, by applying this perturbation to every function in $\mathcal{F}$, we alter relative errors by at most $2\tau/\beta$. By selecting β so that $(1 + 2\tau/\beta)^2 \leq 1 + \varepsilon/2$, we assert that the total error is at most $\varepsilon/2$. To see this, consider any query point x, and let f_i be the function that achieves the minimum value for $\mathcal{F}_{\min}(x)$, and let f'_j be the perturbed function that achieves the minimum value for $\mathcal{F}'_{\min}(x)$. Then

$$\begin{aligned} f_j(x) &\leq \left(1 + \frac{2\tau}{\beta}\right) f'_j(x) \leq \left(1 + \frac{2\tau}{\beta}\right) f'_i(x) \\ &\leq \left(1 + \frac{2\tau}{\beta}\right)^2 f_i(x) \leq \left(1 + \frac{\varepsilon}{2}\right) f_i(x). \end{aligned}$$

It is easy to verify that for all sufficiently small ε, our choice of $\beta = 10\tau/\varepsilon$ satisfies this condition (and it is also at least 2τ as required by the lemma).

We can now explain how to answer ε-ANN queries for the inner cluster. Consider the sites of the inner cluster, which all lie within B_w (see Figure 5(a)). We apply Lemma 4.4 to produce the perturbed family $\mathcal{F}'$ of τ-admissible functions (see Figure 5(b)).



(a) (b) (c)

Figure 5: (a) Inner-cluster sites with their respective distance functions, (b) their perturbation to a common site p', and (c) the reduction to Lemma 3.2.

Since these are all scaling distance functions, the nearest neighbor of any query point $q \in \mathbb{R}^d$ (irrespective of whether it lies within w) is the same for every point on the ray from p' through q. Therefore, it suffices to evaluate the answer to the query for any single query point q' on this ray. In particular, let us fix a hypercube of side length 2 centered at p' (see Figure 5(c)). We will show how to answer $(\varepsilon/3)$-AVR queries for points on the boundary of this hypercube with respect to $\mathcal{F}'$. A general query will then be answered by computing the point where the ray from p' to the query point intersects the hypercube's boundary and returning the result of this query. The total error with respect to the original functions will be at most $(1 + \varepsilon/2)(1 + \varepsilon/3)$, and for all sufficiently small ε, this is at most $1 + \varepsilon$, as desired.

All that remains is to show how to answer $(\varepsilon/3)$-AVR queries for points on the boundary of the hypercube. Let $s = 1/(2\tau + 1)$, and let W be a set of hypercubes of diameter s that cover the

boundary of the hypercube of side length 2 centered at p' (see Figure 5(c)). The number of such boxes is $O(\tau^{d-1})$. For each $w' \in W$, let $B_{w'}$ be the smallest ball enclosing w'. Each point on the hypercube is at distance at least 1 from p'. For each $w' \in W$, we have $\text{dist}(p', B_{w'}) \geq 1 - s = 2\tau \cdot \text{diam}(B_{w'})$, implying that p' and $B_{w'}$ are (2τ)-separated. Therefore, by Lemma 3.2 there is a data structure that can answer $(\varepsilon/3)$-AVR queries with respect to the perturbed distance functions $\mathcal{F}'_{\min}$ in time $O(\log \frac{1}{\varepsilon})$ with storage $O((\frac{1}{\varepsilon})^{d/2})$.

In summary, a query is answered by computing the ray from p' through q, and determining the unique point q' on the boundary of the hypercube that is hit by this ray. We then determine the hypercube w' containing q' in constant time and invoke the associated data structure for answering $(\varepsilon/3)$-AVR queries with respect to $\mathcal{F}'$. The total storage needed for all these structures is $O(\tau^{d-1}/\varepsilon^{d/2})$. For any query point, we can determine which of these data structures to access in $O(1)$ time. Relative to the case of the outer cluster, we suffer only an additional factor of $O(\tau^{d-1})$ to store these data structures.

Under our assumption that γ and σ are constants, it follows that both τ and α are constants and β is $O(1/\varepsilon)$. By Lemma 4.1, the total number of leaf nodes in the (α, β)-AVD is $O(n \log \frac{1}{\varepsilon})$. Combining this with the $O(1/\varepsilon^{d/2})$ space for the data structure to answer queries with respect to the outer cluster and $O(\tau^{d-1}/\varepsilon^{d/2})$ overall space for the inner cluster, we obtain a total space of $O((n \log \frac{1}{\varepsilon})/\varepsilon^{d/2})$. The query time is simply the combination of the $O(\log(\alpha n) + \log\log \beta) = O(\log n + \log\log \frac{1}{\varepsilon})$ time to locate the leaf cell (by Lemma 4.1), and the $O(\log \frac{1}{\varepsilon})$ time to answer $O(\varepsilon)$-AVR queries. The total query time is therefore $O(\log \frac{n}{\varepsilon})$, as desired. This establishes Theorem 1.1.

5 Answering ANN Queries for Bregman Divergences

In this section we demonstrate how to answer ε-ANN queries for a set of n sites over a Bregman divergence. We assume that the Bregman divergence is defined by a strictly convex, twice-differentiable function F over an open convex domain $\mathcal{X} \subseteq \mathbb{R}^d$. As mentioned in the introduction, given a site p, we interpret the divergence $D_F(x, p)$ as a distance function of x about p, that is, analogous to $f_p(x)$ for scaling distances. Thus, gradients and Hessians are defined with respect to the variable x. Our results will be based on the assumption that the divergence is τ-admissible for a constant τ. This will be defined formally in the following section.

5.1 Measures of Bregman Complexity

In Section 1 we introduced the concepts of similarity and asymmetry for Bregman divergences. We can extend the notion of admissibility to Bregman divergences by defining a Bregman divergence D_F to be *τ-admissible* if the associated distance function $f_p(\cdot) = D_F(\cdot, p)$ is τ-admissible.

It is natural to ask how the various criteria of Bregman complexity (asymmetry, similarity, and admissibility) relate to each other. For the sake of relating admissibility with asymmetry, it will be helpful to introduce a directionally-sensitive variant of admissibility. Given f_p and τ as above, we say that f_p is *directionally τ-admissible* if for all $x \in \mathcal{X}$, $\nabla f_p(x) \cdot (x - p) \leq \tau f_p(x)$. (Note that only the gradient condition is used in this definition.) The following lemma provides some comparisons. The proof is rather technical and has been deferred to Section 6.

Lemma 5.1. *Given an open convex domain $\mathcal{X} \subseteq \mathbb{R}^d$:*

(*i*) *Any μ-similar Bregman divergence over $\mathcal{X}$ is 2μ-admissible.*

(*ii*) *Any μ-admissible Bregman divergence over $\mathcal{X}$ is directionally μ-admissible.*

(iii) *A Bregman divergence over $\mathcal{X}$ is μ-asymmetric if and only if it is directionally $(1+\mu)$-admissible.*

Note that claim (i) is strict since the Bregman divergence D_F defined by $F(x) = x^4$ over $\mathcal{X} = \mathbb{R}$ is not μ-similar for any μ, but it is 4-admissible. We do not know whether claim (ii) is strict, but we conjecture that it is.

5.2 ANN Data Structure for Bregman Divergences

Let us return to the discussion of how to answer ε-ANN queries for a τ-admissible Bregman divergence over a domain $\mathcal{X}$. Because any distance function that is τ-admissible is τ'-admissible for any $\tau' \geq \tau$, we may assume that $\tau \geq 1$.[8] We begin by building an (α, β)-AVD over $\mathbb{R}^d$ by invoking Lemma 4.1 for $\alpha = 2\tau$ and $\beta = 4\tau^2/\varepsilon$. (These choices will be justified below.) For each leaf cell w, the nearest neighbor of any query point $q \in w$ can arise from one of the three cases in the lemma. Cases (i) and (ii) are handled in exactly the same manner as in Section 4.3. (Case (i) is trivial, and case (ii) applies for any τ-admissible family of functions.)

It remains to handle case (iii), the *inner cluster*. Recall that these sites lie within a ball B_w such that $\text{dist}(B_w, w) \geq \beta \cdot \text{diam}(B_w)$. We show that as a result of choosing β sufficiently large, for any query point in w the distance from all the sites within B_w are sufficiently close that we may select any of these sites as the approximate nearest neighbor. This is a direct consequence of the following lemma. The proof has been deferred to Section 6.

Lemma 5.2. *Let D be a τ-admissible Bregman divergence and let $0 < \varepsilon \leq 1$. Consider any leaf cell w of the (α, β)-AVD, where $\beta \geq 4\tau^2/\varepsilon$. Then, for any $q \in w$ and points $p, p' \in B_w$*

$$\frac{|D(q,p) - D(q,p')|}{D(q,p)} \;\leq\; \varepsilon.$$

Under our assumption that τ is a constant, α is a constant and β is $O(1/\varepsilon)$. The analysis proceeds much like the case for scaling distances. By Lemma 4.1, the total number of leaf nodes in the (α, β)-AVD is $O(n \log \frac{1}{\varepsilon})$. We require only one representative for cases (i) and (iii), and as in Section 4.3, we need space $O(1/\varepsilon^{d/2})$ to handle case (ii). The query time is simply the combination of the $O(\log(\alpha n) + \log\log\beta) = O(\log n + \log\log\frac{1}{\varepsilon})$ time to locate the leaf cell (by Lemma 4.1), and the $O(\log\frac{1}{\varepsilon})$ time to answer $O(\varepsilon)$-AVR queries for case (ii). The total query time is therefore $O(\log\frac{n}{\varepsilon})$, as desired. This establishes Theorem 1.2.

6 Deferred Technical Details

In this section we present a number of technical results and proofs, which have been deferred from the main presentation.

6.1 On Vertical Ray Shooting

We present a proof of Lemma 2.1 from Section 2.2, which shows how to answer approximate vertical ray-shooting queries for the lower envelope of concave functions in a very restricted context.

[8] Indeed, it can be shown that any distance function that is convex, as Bregman divergences are, cannot be τ-admissible for $\tau < 1$.

Lemma 2.1. (Answering ε-AVR Queries) *Consider a unit ball $B \subseteq \mathbb{R}^d$ and a family of concave functions $\mathcal{F} = \{f_1, \ldots, f_m\}$ defined over B such that for all $1 \leq i \leq m$ and $x \in B$, $f_i(x) \in [0,1]$ and $\|\nabla f_i(x)\| \leq 1$. Then, for any $0 < \varepsilon \leq 1$, there is a data structure that can answer absolute ε-AVR queries in time $O(\log \frac{1}{\varepsilon})$ and storage $O((\frac{1}{\varepsilon})^{d/2})$.*

Proof. We will follow the strategy presented in [8] for answering ε-ANN queries. It combines (1) a data structure for answering approximate central ray-shooting queries, in which the rays originate from a common point and (2) an approximation-preserving reduction from vertical to central ray-shooting queries [6].

Let K denote a closed convex body that is represented as the intersection of a finite set of halfspaces. We assume that K is centrally γ-fat for some constant γ (recall the definition from Section 4.2). An *ε-approximate central ray-shooting query* (ε-ACR query) is given a query ray that emanates from the origin and returns the index of one of K's bounding hyperplanes h whose intersection with the ray is within distance $\varepsilon \cdot \mathrm{diam}(K)$ of the true contact point with K's boundary. We will make use of the following result, which is paraphrased from [6].

Approximate Central Ray-Shooting: Given a convex polytope K in $\mathbb{R}^d$ that is centrally γ-fat for some constant γ and an approximation parameter $0 < \varepsilon \leq 1$, there is a data structure that can answer ε-ACR queries in time $O(\log \frac{1}{\varepsilon})$ and storage $O(1/\varepsilon^{(d-1)/2})$.

As in Section 4 of [6], we can employ a projective transformation that converts vertical ray shooting into central ray shooting. While the specific transformation presented there was tailored to work for a set of hyperplanes that are tangent to a paraboloid, a closer inspection reveals that the reduction can be generalized (with a change in the constant factors) provided that the following quantities are all bounded above by a constant: (1) the diameter of the domain of interest, (2) the difference between the maximum and minimum function values throughout this domain, and (3) the absolute values of the slopes of the hyperplanes (or equivalently, the norms of the gradients of the functions defined by these hyperplanes). This projective transformation produces a convex body in $\mathbb{R}^{d+1}$ that is centrally γ-fat for some constant γ, and it preserves relative errors up to a constant factor.

Therefore, by applying this projective transformation, we can reduce the problem of answering ε-AVR queries in dimension d for the lower envelope of a set of linear functions to the aforementioned ACR data structure in dimension $d+1$. The only remaining issue is that the functions of $\mathcal{F}$ are concave, not necessarily linear. Thus, the output of the reduction is a convex body bounded by curved patches, not a polytope. We address this by applying Dudley's Theorem [23] to produce a polytope that approximates this convex body to an absolute Hausdorff error of $\varepsilon/2$. (In particular, Dudley's construction samples $O(1/\varepsilon^{d/2})$ points on the boundary of the convex body, and forms the approximation by intersecting the supporting hyperplanes at each of these points.) We then apply the ACR data structure to this approximating polytope, but with the allowed error parameter set to $\varepsilon/2$. The combination of the two errors, results in a total allowed error of ε.

In order to obtain a witness, each sample point from Dudley's construction is associated with the function(s) that are incident to that point. We make the general position assumption that no more than $d+1$ functions can coincide at any point on the lower envelope of $\mathcal{F}$, and hence each sample point is associated with a constant number of witnesses. The witness produced by the ACR data structure will be one of the bounding hyperplanes. We check each of the functions associated with the sample point that generated this hyperplane, and return the index of the function having the smallest function value. □

6.2 On Admissibility and Scaling Functions

Next, we present a proof of Lemma 3.1 from Section 3.2. It establishes the key properties of admissible functions, which will be used throughout our analysis.

Lemma 3.1. *Consider an open convex domain $\mathcal{X}$, a site $p \in \mathcal{X}$, a τ-admissible distance function f_p, and a Euclidean ball $B \subset \mathcal{X}$. If B and p are $(\tau\kappa)$-separated for $\kappa > 1$, then:*

(*i*) $f_p^+(B) \leq f_p^-(B)\kappa/(\kappa-1)$,

(*ii*) $\|\nabla f_p^+(B)\| \leq f_p^+(B)/(\kappa \operatorname{diam}(B))$, *and*

(*iii*) $\|\nabla^2 f_p^+(B)\| \leq f_p^+(B)/(\kappa \operatorname{diam}(B))^2$.

Proof. To prove (i), let x^+ and x^- denote the points of B that realize the values of $f_p^+(B)$ and $f_p^-(B)$, respectively. By applying the mean value theorem, there exists a point $s \in \overline{x^- x^+}$ such that $f_p^+(B) - f_p^-(B) = \nabla f_p(s) \cdot (x^+ - x^-)$. By the Cauchy-Schwarz inequality

$$f_p^+(B) - f_p^-(B) = \nabla f_p(s) \cdot (x^+ - x^-) \leq \|\nabla f_p(s)\| \|x^+ - x^-\|.$$

By τ-admissibility, $\|\nabla f_p(s)\| \leq \tau f_p(s)/\|s-p\|$, and since $x^+, x^-, s \in B$, we have $\|x^+ - x^-\|/\|s-p\| \leq \operatorname{diam}(B)/\operatorname{dist}(p,B) \leq 1/(\tau\kappa)$. Thus,

$$f_p^+(B) - f_p^-(B) \leq \frac{\tau f_p(s)}{\|s-p\|}\|x^+ - x^-\| \leq \frac{\tau f_p(s)}{\tau\gamma} \leq \frac{f_p^+(B)}{\kappa}.$$

This implies that $f_p^+(B) \leq f_p^-(B)\kappa/(\kappa-1)$, establishing (i).

To prove (ii), consider any $x \in B$. By separation, $\operatorname{dist}(p,B) \geq \tau\kappa \operatorname{diam}(B)$. Combining this with τ-admissibility and (i), we have

$$\|\nabla f_p(x)\| \leq \frac{\tau f_p(x)}{\|x-p\|} \leq \frac{\tau f_p^+(B)}{\operatorname{dist}(p,B)} \leq \frac{\tau f_p^+(B)}{\tau\kappa \operatorname{diam}(B)} = \frac{f_p^+(B)}{\kappa \operatorname{diam}(B)}.$$

This applies to any $x \in B$, thus establishing (ii).

To prove (iii), again consider any $x \in B$. By separation and admissibility, we have

$$\|\nabla^2 f_p(x)\| \leq \frac{\tau^2 f_p(x)}{\|x-p\|^2} \leq \frac{\tau^2 f_p^+(B)}{\operatorname{dist}^2(p,B)} \leq \frac{f_p^+(B)}{(\kappa \operatorname{diam}(B))^2}.$$

This applies to any $x \in B$, thus establishing (iii). □

We next present a proof of the three properties of the normalized distance functions that were introduced in the proof of Lemma 3.2. These properties follow in a straightforward manner from admissibility.

Lemma 6.1. *Each of the normalized distance functions $g(x) = f(x')/h$ defined in the proof of Lemma 3.2 satisfy the following properties:*

(a) $g^+(B_0) \leq 4/5$ *and* $g^-(B_0) \geq 1/5$,

(b) $\|\nabla g^+(B_0)\| \leq 1/2$, *and*

(c) $\|\nabla^2 g^+(B_0)\| \leq 1/4$.

Proof. For any $x \in B_0$, we have

$$g(x) \;\leq\; \frac{f^+(B)}{h} \;\leq\; \frac{2f^-(B)}{h} \;\leq\; \frac{4f_1^-(B)}{h} \;=\; \frac{4}{5},$$

and

$$g(x) \;\geq\; \frac{f^-(B)}{h} \;\geq\; \frac{f_1^-(B)}{h} \;=\; \frac{1}{5},$$

which establishes (a).

Before establishing (b) and (c), observe that by the chain rule in differential calculus, $\nabla g(x) = (r/h)\nabla f(x')$ and $\nabla^2 g(x) = (r^2/h)\nabla^2 f(x')$. (Recall that x and x' are corresponding points in B_0 and B, respectively.) Since B_0 is a unit ball, $\mathrm{diam}(B_0) = 2$. Thus, by Corollary 3.1(ii),

$$\|\nabla g(x)\| \;=\; \frac{r}{h}\|\nabla f(x')\| \;\leq\; \frac{r}{h}\frac{f^+(B)}{2(2r)} \;\leq\; \frac{1}{4},$$

which establishes (b). By Corollary 3.1(iii),

$$\|\nabla^2 g(x)\| \;=\; \frac{r^2}{h}\|\nabla f(x')\| \;\leq\; \frac{r^2}{h}\frac{f^+(B)}{(2(2r))^2} \;\leq\; \frac{1}{16},$$

which establishes (c). □

Here is a proof of Lemma 4.2 from Section 4.2, which relates the admissibility of a scaling distance function to the fatness and smoothness of the associated metric ball.

Lemma 4.2. *Given positive reals γ and σ, let f_p be a (γ, σ)-distance over $\mathbb{R}^d$ scaled about some point $p \in \mathbb{R}^d$. There exists τ (a function of γ and σ) such that f_p is τ-admissible.*

Proof. For any point $x \in \mathbb{R}^d$, we will show that (i) $\|\nabla f_p(x)\| \cdot \|x - p\| \leq f_p(x)/\gamma$ and (ii) $\|\nabla^2 f_p(x)\| \cdot \|x - p\|^2 \leq 2f_p(x)/(\sigma\gamma^3)$. It will follow that f_p is τ-admissible for $\tau = \sqrt{2/(\sigma\gamma^3)}$.

Let K denote the unit metric ball associated with f_p and let K' denote the scaled copy of K that just touches the point x. Let r be the unit vector in the direction px (we refer to this as the *radial direction*), and let n be the outward unit normal vector to the boundary of K' at x. (Throughout the proof, unit length vectors are defined in the Euclidean sense.) As K' is centrally γ-fat, it is easy to see that the cosine of the angle between r and n, that is, $r \cdot n$, is at least γ. As the boundary of K' is the level surface of f_p, it follows that $\nabla f_p(x)$ is directed along n. To compute the norm of the gradient, note that

$$\nabla f_p(x) \cdot r \;=\; \lim_{\delta \to 0} \frac{f_p(x + \delta r) - f_p(x)}{\delta}.$$

As f_p is a scaling distance function, it follows that

$$f_p(x + \delta r) - f_p(x) \;=\; \frac{\delta}{\|x - p\|} f_p(x).$$

Thus

$$\nabla f_p(x) \cdot r \;=\; \frac{f_p(x)}{\|x - p\|}.$$

Recalling that $r \cdot n \geq \gamma$, we obtain

$$\|\nabla f_p(x)\| \;\leq\; \frac{f_p(x)}{\gamma\|x - p\|}.$$

Thus $\|\nabla f_p(x)\| \cdot \|x - p\| \leq f_p(x)/\gamma$, as desired.

We next bound the norm of the Hessian $\nabla^2 f_p(x)$. As the Hessian matrix is positive semidefinite, recall that it has a full set of independent eigenvectors that are mutually orthogonal, and its norm equals its largest eigenvalue. Because f_p is a scaling distance function, it changes linearly along the radial direction. Therefore, one of the eigenvectors of $\nabla^2 f_p(x)$ is in direction r, and the associated eigenvalue is 0 (see Figure 6). It follows that the remaining eigenvectors all lie in a subspace that is orthogonal to r. In particular, the eigenvector associated with its largest eigenvalue must lie in this subspace. Let u denote such an eigenvector of unit length, and let λ denote the associated eigenvalue.

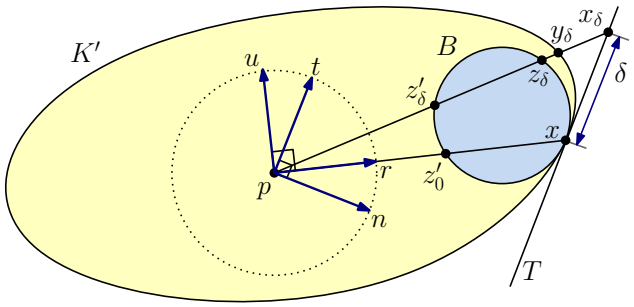


Figure 6: Proof of Lemma 4.2.

Note that λ is the second directional derivative of f_p in the direction u. In order to bound λ, we find it convenient to first bound the second directional derivative of f_p in a slightly different direction. Let T denote the hyperplane tangent to K' at point x. We project u onto T and let t denote the resulting vector scaled to have unit length. We will compute the second directional derivative of f_p in the direction t. Let λ_t denote this quantity. In order to relate λ_t with λ, we write t as $(t \cdot r)r + (t \cdot u)u$. Since r and u are mutually orthogonal eigenvectors of $\nabla^2 f_p(x)$, by elementary linear algebra, it follows that $\lambda_t = (t \cdot r)^2 \lambda_r + (t \cdot u)^2 \lambda_u$, where λ_r and λ_u are the eigenvalues associated with r and u, respectively. Since $\lambda_r = 0$, $\lambda_u = \lambda$, and $t \cdot u = r \cdot n \geq \gamma$, we have $\lambda_t \geq \gamma^2 \lambda$, or equivalently, $\lambda \leq \lambda_t/\gamma^2$. In the remainder of the proof, we will bound λ_t, which will yield the desired bound on λ.

Let $x_\delta = x + \delta t$ and $\psi(\delta) = f_p(x_\delta)$. Clearly $\lambda_t = \psi''(0)$. Using the Taylor series and the fact that $\psi'(0) = \nabla f_p(x) \cdot t = 0$, it is easy to see that

$$\psi''(0) = 2 \cdot \lim_{\delta \to 0} \frac{\psi(\delta) - \psi(0)}{\delta^2}.$$

Letting y_δ denote the intersection point of the segment $\overline{px_\delta}$ with the boundary of K', and observing that both x and y_δ lie on $\partial K'$ (implying that $f_p(x) = f_p(y_\delta)$), we have

$$\psi(\delta) \;=\; f_p(x_\delta) \;=\; \frac{\|x_\delta - p\|}{\|y_\delta - p\|} f_p(x),$$

and thus

$$\psi(\delta) - \psi(0) \;=\; \frac{\|x_\delta - p\| - \|y_\delta - p\|}{\|y_\delta - p\|} f_p(x) \;=\; \frac{\|x_\delta - y_\delta\|}{\|y_\delta - p\|} f_p(x).$$

It follows that

$$\psi''(0) \;=\; 2 \cdot \lim_{\delta \to 0} \frac{1}{\delta^2} \frac{\|x_\delta - y_\delta\|}{\|y_\delta - p\|} f_p(x) \;=\; \frac{2 f_p(x)}{\|x - p\|} \cdot \lim_{\delta \to 0} \frac{\|x_\delta - y_\delta\|}{\delta^2}.$$

We next compute this limit. Let $B \subset K'$ denote the maximal ball tangent to K' at x and let R denote its radius. As K' is σ-smooth, we have that

$$R \;\geq\; \frac{\sigma}{2} \cdot \mathrm{diam}(K') \;\geq\; \frac{\sigma}{2} \cdot \|x - p\|.$$

Consider the line passing through p and x_δ. For sufficiently small δ, it is clear that this line must intersect the boundary of the ball B at two points. Let z_δ denote the intersection point closer to x_δ and z'_δ denote the other intersection point. Clearly, $\|x_\delta - y_\delta\| \leq \|x_\delta - z_\delta\|$ and, by the power of the point theorem, we have

$$\delta^2 \;=\; \|x_\delta - x\|^2 \;=\; \|x_\delta - z_\delta\| \cdot \|x_\delta - z'_\delta\|.$$

It follows that

$$\frac{\|x_\delta - y_\delta\|}{\delta^2} \;\leq\; \frac{\|x_\delta - z_\delta\|}{\delta^2} \;=\; \frac{1}{\|x_\delta - z'_\delta\|}.$$

Thus

$$\lim_{\delta \to 0} \frac{\|x_\delta - y_\delta\|}{\delta^2} \;\leq\; \lim_{\delta \to 0} \frac{1}{\|x_\delta - z'_\delta\|} \;=\; \frac{1}{\|x - z'_0\|},$$

where z'_0 denotes the point of intersection of the line passing through p and x with the boundary of B. Since the cosine of the angle between this line and the diameter of ball B at x equals $r \cdot n$, which is at least γ, we have $\|x - z'_0\| \geq 2\gamma R$. It follows that

$$\lim_{\delta \to 0} \frac{\|x_\delta - y_\delta\|}{\delta^2} \;\leq\; \frac{1}{2\gamma R} \;\leq\; \frac{1}{\sigma\gamma\|x - p\|}.$$

Substituting this bound into the expression found above for λ_t, we obtain

$$\lambda_t \;=\; \psi''(0) \;\leq\; \frac{2f_p(x)}{\sigma\gamma\|x - p\|^2}.$$

Recalling that $\lambda \leq \lambda_t/\gamma^2$, we have

$$\lambda \;\leq\; \frac{2f_p(x)}{\sigma\gamma^3\|x - p\|^2},$$

which implies that $\|\nabla^2 f_p(x)\| \cdot \|x - p\|^2 \leq 2f_p(x)/(\sigma\gamma^3)$. This completes the proof. □

6.3 On Bregman Divergences

The following lemma provides some properties of Bregman divergences, which will be used later. Throughout, we assume that a Bregman divergence is defined by a strictly convex, twice-differentiable function F over an open convex domain $\mathcal{X} \subseteq \mathbb{R}^d$. Given a site p, we interpret the divergence $D_F(x, p)$ as a distance function of x about p, and so gradients and Hessians are defined with respect to the variable x. The following lemma provides a few useful observations regarding the Bregman divergence. We omit the proof since these all follow directly from the definition of Bregman divergence. Observation (i) is related to the *symmetrized Bregman divergence* [1]. Observation (ii), known as the *three-point property* [16], generalizes the law of cosines when the Bregman divergence is the Euclidean squared distance.

Lemma 6.2. *Given any Bregman divergence D_F defined over an open convex domain $\mathcal{X}$, and points $q, p, p' \in \mathcal{X}$:*

(*i*) $D_F(q, p) + D_F(p, q) = (\nabla F(q) - \nabla F(p)) \cdot (q - p)$

(*ii*) $D_F(q,p') + D_F(p',p) = D_F(q,p) + (q-p') \cdot (\nabla F(p) - \nabla F(p'))$

(*iii*) $\nabla D_F(q,p) = \nabla F(q) - \nabla F(p)$

(*iv*) $\nabla^2 D_F(q,p) = \nabla^2 F(q)$.

In parts (iii) and (iv), derivatives involving $D_F(q,p)$ are taken with respect to q.

The above result allows us to establish the following upper and lower bounds on the value, gradient, and Hessian of a Bregman divergence based on the maximum and minimum eigenvalues of the function's Hessian.

Lemma 6.3. *Let F be a strictly convex function defined over some domain $\mathcal{X} \subseteq \mathbb{R}^d$, and let D_F denote the associated Bregman divergence. For each $x \in \mathcal{X}$, let $\lambda_{\min}(x)$ and $\lambda_{\max}(x)$ denote the minimum and maximum eigenvalues of $\nabla^2 F(x)$, respectively. Then, for all $p, q \in \mathcal{X}$, there exist points r_1, r_2, and r_3 on the open line segment $\overline{pq}$ such that*

$$\begin{aligned} \tfrac{1}{2}\lambda_{\min}(r_1)\|q-p\|^2 &\leq D_F(q,p) \leq \tfrac{1}{2}\lambda_{\max}(r_1)\|q-p\|^2 \\ \lambda_{\min}(r_2)\|q-p\| &\leq \|\nabla D_F(q,p)\| \leq \lambda_{\max}(r_3)\|q-p\| \\ \lambda_{\min}(q) &\leq \|\nabla^2 D_F(q,p)\| \leq \lambda_{\max}(q). \end{aligned}$$

Proof. To establish the first inequality, we apply Taylor's theorem with the Lagrange form of the remainder to obtain

$$F(q) = F(p) + \nabla F(p) \cdot (q-p) + \frac{1}{2}(q-p)^\intercal \nabla^2 F(r_1)(q-p),$$

for some r_1 on the open line segment $\overline{pq}$. By substituting the above expression for $F(q)$ into the definition of $D_F(q,p)$ we obtain

$$D_F(q,p) = F(q) - F(p) - \nabla F(p) \cdot (q-p) = \frac{1}{2}(q-p)^\intercal \nabla^2 F(r_1)(q-p).$$

By basic linear algebra, we have

$$\lambda_{\min}(r_1)\|q-p\|^2 \leq (q-p)^\intercal \nabla^2 F(r_1)(q-p) \leq \lambda_{\max}(r_1)\|q-p\|^2.$$

Therefore,

$$\frac{\lambda_{\min}(r_1)}{2}\|q-p\|^2 \leq D_F(q,p) \leq \frac{\lambda_{\max}(r_1)}{2}\|q-p\|^2,$$

which establishes the first assertion.

For the second assertion, we recall from Lemma 6.2(iii) that $\nabla D_F(q,p) = \nabla F(q) - \nabla F(p)$. Let v be any unit vector. By applying the mean value theorem to the function $\psi(t) = v^\intercal \nabla F(p + t(q-p))$ for $0 \leq t \leq 1$, there exists a point $r_2 \in \overline{pq}$ (which depends on v) such that $v^\intercal(\nabla F(q) - \nabla F(p)) = v^\intercal \nabla^2 F(r_2)(q-p)$. Taking v to be the unit vector in the direction of $q-p$, and applying the Cauchy-Schwarz inequality, we obtain

$$\|\nabla F(q) - \nabla F(p)\| \geq |v^\intercal(\nabla F(q) - \nabla F(p))| = |v^\intercal \nabla^2 F(r_2)(q-p)| \geq \lambda_{\min}(r_2)\|q-p\|.$$

For the upper bound, we apply the same approach, but take v to be the unit vector in the direction of $\nabla F(q) - \nabla F(p)$. There exists $r_3 \in \overline{pq}$ such that

$$\|\nabla F(q) - \nabla F(p)\| = |v^\intercal(\nabla F(q) - \nabla F(p))| = |v^\intercal \nabla^2 F(r_3)(q-p)| \leq \|\nabla^2 F(r_3)(q-p)\| \leq \lambda_{\max}(r_3)\|q-p\|.$$

This establishes the second assertion.

The final assertion follows from the fact that $\nabla^2 D_F(q,p) = \nabla^2 F(q)$ (Lemma 6.2(iv)) and the definition of the spectral norm. □

With the help of this lemma, we can now present a proof of Lemma 5.1 from Section 5.1, which relates the various measures of complexity for Bregman divergences.

Lemma 5.1. *Given an open convex domain $\mathcal{X} \subseteq \mathbb{R}^d$:*

(*i*) *Any μ-similar Bregman divergence over $\mathcal{X}$ is 2μ-admissible.*

(*ii*) *Any μ-admissible Bregman divergence over $\mathcal{X}$ is directionally μ-admissible.*

(*iii*) *A Bregman divergence over $\mathcal{X}$ is μ-asymmetric if and only if it is directionally $(1+\mu)$-admissible.*

Proof. For each $x \in \mathcal{X}$, let $\lambda_{\min}(x)$ and $\lambda_{\max}(x)$ denote the minimum and maximum eigenvalues of $\nabla^2 F(x)$, respectively. We first show that for all $x \in \mathcal{X}$, $2 \leq \lambda_{\min}(x)$ and $\lambda_{\max}(x) \leq 2\mu$. We will prove only the second inequality, since the first follows by a symmetrical argument. Suppose to the contrary that there was a point $x \in \mathcal{X}$ such that $\lambda_{\max}(x) > 2\mu$. By continuity and the fact that $\mathcal{X}$ is convex and open, there exists a point $q \in \mathcal{X}$ distinct from x such that for any r on the open line segment $\overline{qx}$,

$$(q-x)^\intercal \nabla^2 F(r)(q-x) \;>\; 2\mu \|q-x\|^2. \tag{1}$$

Specifically, we may take q to lie sufficiently close to x along $x+v$, where v is the eigenvector associated with $\lambda_{\max}(x)$. As in the proof of Lemma 6.3, we apply Taylor's theorem with the Lagrange form of the remainder to obtain

$$\begin{aligned} D_F(q,x) \;&=\; F(q) - F(x) - \nabla F(x)\cdot(q-x) \\ &=\; \frac{1}{2}(q-x)^\intercal \nabla^2 F(r)(q-x) \;=\; \left(\frac{1}{t}\right)^2 \frac{1}{2}(r-x)^\intercal \nabla^2 F(r)(r-x). \end{aligned}$$

By Eq. (1), we have $D_F(q,x) > \mu\|q-x\|^2$. Therefore, D_F is not μ-similar. This yields the desired contradiction.

Because $2 \leq \lambda_{\min}(x) \leq \lambda_{\max}(x) \leq 2\mu$ for all $x \in \mathcal{X}$, by Lemma 6.3, we have

$$\|q-p\|^2 \;\leq\; D_F(q,p), \qquad \|\nabla D_F(q,p)\| \;\leq\; 2\mu\|q-p\|, \quad \text{and} \quad \|\nabla^2 D_F(q,p)\| \;\leq\; 2\mu,$$

which imply

$$\|\nabla D_F(q,p)\|\,\|q-p\| \;\leq\; 2\mu D_F(q,p) \qquad \text{and} \qquad \|\nabla^2 D_F(q,p)\|\,\|q-p\|^2 \;\leq\; 2\mu D_F(q,p),$$

which together imply that D is 2μ-admissible, as desired.

To prove (ii), observe that by the Cauchy-Schwarz inequality $\nabla D_F(q,p)\cdot(q-p) \leq \|\nabla D_F(q,p)\| \cdot \|q-p\|$, and therefore, any divergence that satisfies the condition for μ-admissibility immediately satisfies the condition for directional μ-admissibility.

To show (iii), consider any points $p, q \in \mathcal{X}$. Recall the facts regarding the Bregman divergence presented in Lemma 6.2. By combining observations (i) and (iii) from that lemma, we have $D_F(q,p) + D_F(p,q) = \nabla D_F(q,p)\cdot(q-p)$. Observe that if D is directionally $(1+\mu)$-admissible, then

$$D_F(q,p) + D_F(p,q) \;=\; \nabla D_F(q,p)\cdot(q-p) \;\leq\; (1+\mu)D_F(q,p),$$

which implies that $D_F(p,q) \leq \mu(D_F(q,p)$, and hence D is μ-asymmetric. Conversely, if D is μ-asymmetric, then

$$\nabla D_F(q,p)\cdot(q-p) \;=\; D_F(q,p) + D_F(p,q) \;\leq\; D_F(q,p) + \mu D_F(q,p) \;=\; (1+\mu)D_F(q,p),$$

implying that D_F is directionally $(1+\mu)$-admissible. (Recall that directional admissibility requires only that the gradient condition be satisfied.) □

Next, we provide a proof of Lemma 5.2 from Section 5.2,

Lemma 5.2. *Let D be a τ-admissible Bregman divergence and let $0 < \varepsilon \leq 1$. Consider any leaf cell w of the (α, β)-AVD, where $\beta \geq 4\tau^2/\varepsilon$. Then, for any $q \in w$ and points $p, p' \in B_w$*

$$\frac{|D(q,p) - D(q,p')|}{D(q,p)} \;\leq\; \varepsilon.$$

Proof. Without loss of generality, we may assume that $D(q,p) \geq D(q,p')$. By adding $D(p,p')$ to the left side of Lemma 6.2(ii) and rearranging terms, we have

$$\begin{aligned} D(q,p) - D(q,p') \leq\; & (D(q,p) - D(q,p')) + D(p,p') \\ =\; & (D(p',p) + (\nabla F(p') - \nabla F(p)) \cdot (q - p')) + D(p,p') \\ =\; & (\nabla F(p') - \nabla F(p)) \cdot (q - p')) + (D(p',p) + D(p,p')). \end{aligned}$$

By Lemma 6.2(i) we have

$$\begin{aligned} D(q,p) - D(q,p') \;\leq\; & (\nabla F(p') - \nabla F(p)) \cdot (q - p') + (\nabla F(p') - \nabla F(p)) \cdot (p' - p) \\ =\; & (\nabla F(p') - \nabla F(p)) \cdot (q - p). \end{aligned}$$

Let v be any unit vector. Applying the mean value theorem to the function $\psi(t) = v^\intercal \nabla F(p + t(p' - p))$ for $0 \leq t \leq 1$, implies that there exists a point $r \in \overline{pp'}$ (which depends on v) such that $v^\intercal(\nabla F(p') - \nabla F(p)) = v^\intercal \nabla^2 F(r)(p' - p)$. Taking v to be the unit vector in the direction of $q - p$, and applying the Cauchy-Schwarz inequality, we obtain

$$D(q,p) - D(q,p') \;\leq\; (\nabla^2 F(r)(p' - p)) \cdot (q - p) \;\leq\; \|\nabla^2 F(r)\|\, \|p' - p\|\, \|q - p\|.$$

By Lemma 6.2(iv) and τ-admissibility, $\|\nabla^2 F(r)\| = \|\nabla^2 D(r,q)\| \leq \tau D(r,q)/\|r - q\|^2$, which implies

$$D(q,p) - D(q,p') \;\leq\; \frac{\tau D(r,q)}{\|r - q\|^2}\, \|p' - p\|\, \|q - p\|. \tag{2}$$

Since r lies on the segment between p' and p, it follows that $r \in B_w$. Letting $\delta = \mathrm{diam}(B_w)$, we have $\max(\|p' - p\|, \|r - p\|) \leq \delta$ and $\|r - q\| \geq \beta\delta$. By the triangle inequality, $\|q - p\| \leq \|q - r\| + \|r - p\|$. Therefore,

$$\frac{\|q - p\|}{\|r - q\|} \;\leq\; \frac{\|q - r\| + \|r - p\|}{\|r - q\|} \;=\; 1 + \frac{\|r - p\|}{\|r - q\|} \;\leq\; 1 + \frac{1}{\beta},$$

and since clearly $\beta \geq 1$,

$$\frac{\|p' - p\|\, \|q - p\|}{\|r - q\|^2} \;\leq\; \frac{1}{\beta}\left(1 + \frac{1}{\beta}\right) \;\leq\; \frac{2}{\beta}. \tag{3}$$

We would like to express the right-hand side of Eq. (2) in terms of p rather than r. By the τ-admissibility of D and the fact that $r, p \in B_w$, we can apply Lemma 3.1(i) (with the distance function $f_q(\cdot) = D(\cdot, q)$ and $\kappa = \beta/\tau$) to obtain $D(r,q) \leq D(p,q)/(1 - \tau/\beta)$. Combining Eq. (3) with this, we obtain

$$D(q,p) - D(q,p') \;\leq\; \frac{2\tau}{\beta} D(r,q) \;\leq\; \frac{2\tau}{\beta(1 - \tau/\beta)} D(p,q).$$

In Lemma 5.1(iii) we showed that any $(1 + \mu)$-admissible Bregman divergence is μ-asymmetric, and by setting $\mu = \tau - 1$ it follows that $D(p,q) \leq (\tau - 1) D(q,p)$. Putting this all together, we obtain

$$D(q,p) - D(q,p') \;\leq\; \frac{2\tau(\tau - 1)}{\beta(1 - \tau/\beta)} D(q,p).$$

All that remains is to set β sufficiently large to obtain the desired result. Since $\tau \geq 1$ and $\varepsilon \leq 1$, it is easily verified that setting $\beta = 4\tau^2/\varepsilon$ suffices to produce the desired conclusion. □

Lemma 6.4. *Consider a Bregman divergence D_F defined over an open, convex domain $\mathcal{X}$. Given any invertible affine transformation A, let $G = F \circ A^{-1}$, and let D_G denote the associated Bregman divergence over the domain $\mathcal{Y} = A(\mathcal{X})$. Then for any $q, p \in \mathcal{X}$, $D_F(q,p) = D_G(Aq, Ap)$.*

Proof. By applying the multivariate form of the chain rule with respect to y,the gradient G is

$$\nabla G(y) \;=\; A^{-\intercal} \nabla F(A^{-1}y),$$

where $A^{-\intercal}$ is a shorthand for $(A^{-1})^{\intercal}$.

Let $q' = Aq$ and $p' = Ap$. By definition of the Bregman divergence, we have

$$\begin{aligned} D_G(q',p') &= G(q') - G(p') - \nabla G(p') \cdot (q' - p') \\ &= G(Aq) - G(Ap) - \nabla G(Ap) \cdot (Aq - Ap) \\ &= F(q) - F(p) - (A^{-\intercal} \nabla F(p)) \cdot (A(q-p)). \end{aligned}$$

Using the facts that $u \cdot v = u^{\intercal} v$ and $(uv)^{\intercal} = v^{\intercal} u^{\intercal}$, $(A^{-\intercal} \nabla F(p)) \cdot (A(q-p))$ is equal to

$$\begin{aligned} &= (A^{-\intercal} \nabla F(p))^{\intercal} (A(q-p)) \\ &= \nabla F(p)^{\intercal} (A^{-\intercal})^{\intercal} A(q-p) \\ &= \nabla F(p)^{\intercal} A^{-1} A(q-p) \;=\; \nabla F(p)^{\intercal} (q-p) \\ &= \nabla F(p) \cdot (q-p). \end{aligned}$$

Therefore, $D_G(q',p') = F(q) - F(p) - \nabla F(p) \cdot (q-p) = D_F(q,p)$, as desired. □

References

[1] A. Abdullah, J. Moeller, and S. Venkatasubramanian. Approximate Bregman near neighbors in sublinear time: Beyond the triangle inequality. In *Proc. 28th Annu. Sympos. Comput. Geom.*, pages 31–40, 2012.

[2] A. Abdullah and S. Venkatasubramanian. A directed isoperimetric inequality with application to Bregman near neighbor lower bounds. In *Proc. 47th Annu. ACM Sympos. Theory Comput.*, pages 509–518, 2015.

[3] M. R. Ackermann and J. Blömer. Coresets and approximate clustering for Bregman divergences. In *Proc. 20th Annu. ACM-SIAM Sympos. Discrete Algorithms*, pages 1088–1097, 2009.

[4] I. P. Androulakis, C. D. Maranas, and C. A. Floudas. αBB: A global optimization method for general constrained nonconvex problems. *J. Global Optim.*, 7:337–363, 1995.

[5] S. Arya, G. D. da Fonseca, and D. M. Mount. Near-optimal ε-kernel construction and related problems. In *Proc. 33rd Internat. Sympos. Comput. Geom.*, pages 10:1–15, 2017.

[6] S. Arya, G. D. da Fonseca, and D. M. Mount. Optimal approximate polytope membership. In *Proc. 28th Annu. ACM-SIAM Sympos. Discrete Algorithms*, pages 270–288, 2017.

[7] S. Arya, G. D. da Fonseca, and D. M. Mount. Approximate convex intersection detection with applications to width and Minkowski sums. In *Proc. 26th Annu. European Sympos. Algorithms*, pages 3:1–14, 2018.

[8] S. Arya, G. D. da Fonseca, and D. M. Mount. Approximate polytope membership queries. *SIAM J. Comput.*, 47(1):1–51, 2018.

[9] S. Arya, T. Malamatos, and D. M. Mount. The effect of corners on the complexity of approximate range searching. *Discrete Comput. Geom.*, 41:398–443, 2009.

[10] S. Arya, T. Malamatos, and D. M. Mount. Space-time tradeoffs for approximate nearest neighbor searching. *J. Assoc. Comput. Mach.*, 57:1–54, 2009.

[11] S. Arya and D. M. Mount. Approximate range searching. *Comput. Geom. Theory Appl.*, 17:135–163, 2000.

[12] F. Aurenhammer. Power diagrams: Properties, algorithms and applications. *SIAM J. Comput.*, 16:78–96, 1987.

[13] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *J. of Machine Learning Research*, 6:1705–1749, 2005.

[14] D. P. Bertsekas. Convexification procedures and decomposition methods for nonconvex optimization problems 1. *J. Optim. Theory Appl.*, 29:169–197, 1979.

[15] J.-D. Boissonnat and M. Karavelas. On the combinatorial complexity of Euclidean Voronoi cells and convex hulls of d-dimensional spheres. In *Proc. 14th Annu. ACM-SIAM Sympos. Discrete Algorithms*, pages 305–312, 2003.

[16] J.-D. Boissonnat, F. Nielsen, and R. Nock. Bregman Voronoi diagrams. *Discrete Comput. Geom.*, 44:281–307, 2010.

[17] L. Cayton. Fast nearest neighbor retrieval for Bregman divergences. In *Proc. 25th Internat. Conf. Machine Learning*, pages 112–119, 2008.

[18] T. M. Chan. Applications of Chebyshev polynomials to low-dimensional computational geometry. *J. Comput. Geom.*, 9(2):3–20, 2017.

[19] L. P. Chew and R. L. D. III. Voronoi diagrams based on convex distance functions. In *Proc. First Annu. Sympos. Comput. Geom.*, pages 235–244, 1985.

[20] K. L. Clarkson. A randomized algorithm for closest-point queries. *SIAM J. Comput.*, 17(4):830–847, 1988.

[21] K. L. Clarkson. Building triangulations using ε-nets. In *Proc. 38th Annu. ACM Sympos. Theory Comput.*, pages 326–335, 2006.

[22] M. de Berg, O. Cheong, M. van Kreveld, and M. Overmars. *Computational Geometry: Algorithms and Applications*. Springer, 3rd edition, 2010.

[23] R. M. Dudley. Metric entropy of some classes of sets with differentiable boundaries. *J. Approx. Theory*, 10(3):227–236, 1974.

[24] S. Har-Peled. A replacement for Voronoi diagrams of near linear size. In *Proc. 42nd Annu. IEEE Sympos. Found. Comput. Sci.*, pages 94–103, 2001.

[25] S. Har-Peled and N. Kumar. Approximating minimization diagrams and generalized proximity search. *SIAM J. Comput.*, 44:944–974, 2015.

[26] F. Itakura and S. Saito. Analysis synthesis telephony based on the maximum likelihood method. In *Proc. Sixth Internat. Congress Accoustics*, volume 17, pages C17–C20, 1968.

[27] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22(1):79–86, 1951.

[28] F. Nielsen, P. Piro, and M. Barlaud. Bregman vantage point trees for efficient nearest neighbor queries. *2009 IEEE Internat. Conf. on Multimedia and Expo*, pages 878–881, 2009.

[29] M. Sharir. Almost tight upper bounds for lower envelopes in higher dimensions. *Discrete Comput. Geom.*, 12:327–345, 1994.

[30] S. Si, D. Tao, and B. Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE Trans. Knowl. and Data Eng.*, 22(7):929–942, 2010.