Multiscale Dictionary Learning: Non-Asymptotic Bounds and Robustness

Mauro Maggioni MAURO@MATH.DUKE.EDU
Departments of Mathematics, Electrical and Computer Engineering, and Computer Science
Duke University
Durham, NC 27708, USA

Stanislav Minsker SMINSKER@MATH.DUKE.EDU
Department of Mathematics
Duke University
Durham, NC 27708, USA

Nate Strawn NSTRAWN@MATH.DUKE.EDU
Departments of Mathematics and Electrical and Computer Engineering
Duke University
Durham, NC 27708, USA

Editor:

Abstract

High-dimensional data sets often exhibit inherently low-dimensional structure. Over the past decade, this empirical fact has motivated researchers to study the detection, measurement, and exploitation of such low-dimensional structure, as well as numerous implications for high-dimensional statistics, machine learning, and signal processing. Manifold learning (where the low-dimensional structure is a manifold) and dictionary learning (where the low-dimensional structure is the set of sparse linear combinations of vectors from a finite dictionary) are two prominent theoretical and computational frameworks in this area and, despite their ostensible distinction, the recently-introduced Geometric Multi-Resolution Analysis (GMRA) provides a robust, computationally efficient, multiscale procedure for simultaneously learning a manifold and a dictionary. In this work, we prove non-asymptotic probabilistic bounds on the approximation error of GMRA for a rich class of underlying models that includes "noisy" manifolds, thus theoretically establishing the robustness of the procedure and confirming empirical observations. In particular, if the data aggregates near a low-dimensional manifold, our results show that the approximation error primarily depends on the intrinsic dimension of the manifold, and is independent of the ambient dimension. Our work thus establishes GMRA as a provably fast algorithm for dictionary learning with approximation and sparsity guarantees. We perform numerical experiments that further confirm our theoretical results.

Keywords: Dictionary learning, Multi-Resolution Analysis, Manifold Learning, Robustness, Sparsity

Contents

1. Introduction

In many high-dimensional data analysis problems, existence of *efficient data representations* can dramatically boost the statistical performance and the computational efficiency of learning algorithms. Inversely, in the absence of efficient representations, the curse of dimensionality implies that the required sample size must grow exponentially in the ambient dimension, which renders many statistical learning tasks completely untenable. Parametric statistical modeling seeks to resolve this difficulty by restricting the family of candidate distributions for the data to a collection of probability measures indexed by a finite-dimensional parameter. By contrast, nonparametric statistical models are more flexible and oftentimes more precise, but usually require data samples of large sizes unless the data exhibits some simple latent structure (e.g., some form of sparsity). Such structural considerations are essential for establishing convergence rates, and oftentimes these structural considerations are geometric in nature.

One classical geometric assumption asserts that the data, modeled as a set of points in $\mathbb{R}^D$, in fact lies on (or perhaps very close to) *a single d-dimensional affine subspace* $V \in \mathbb{R}^D$ where $d \ll D$. Tools such as PCA (see Hotelling, 1933, 1936; Pearson, 1901) estimate V in a stable fashion under suitable assumptions. Generalizing this model, one may assert that the data lies on a union of several low-dimensional affine subspaces instead of just one, and in this case the estimation of the *multiple affine subspaces* from data samples already inspired intensive research due to its subtle complexity (e.g., see Chen and Maggioni, 2011; Chen and Lerman, 2009; Elhamifar and Vidal, 2009; Fischler and Bolles, 1981; Ho et al., 2003; Liu et al., 2010; Ma et al., 2007, 2008; Sugaya and Kanatani, 2004; Tipping and Bishop, 1999;

Vidal et al., 2005; Yan and Pollefeys, 2006; Zhang et al., 2010). A widely used form of this model is that of k-sparse data, where there exists a dictionary (i.e., a collection of vectors) $\Phi = \{\varphi_i\}_{i=1}^m \subset \mathbb{R}^D$ such that each observed data point $x \in \mathbb{R}^d$ may be expressed as a linear combination of at most $k \ll D$ elements of Φ. These *sparse representations* offer great convenience and expressivity for signal processing tasks (such as in Peyré, 2009; Protter and Elad, 2007), compressive sensing, statistical estimation ,and learning (e.g., see Aharon et al., 2005; Candes and Tao, 2007; Chen et al., 1998; Donoho, 2006; Kreutz-Delgado et al., 2003; Lewicki et al., 1998; Maurer and Pontil, 2010a, among others), and even exhibits connections with representations in the visual cortex (see Olshausen and Field, 1997). In geometric terminology, such sparse representations are generally attainable when the local *intrinsic dimension* of the observations is small. For these applications, the dictionary is usually assumed to be known a priori, instead of being learned from the data, but it has been recognized in the past decade that data-dependent dictionaries may perform significantly better than generic dictionaries even in classical signal processing tasks.

The k-sparse data model motivates a large amount of research in dictionary learning, where Φ is learned from data rather than being fixed in advance: given n samples $X_1, \ldots, X_n$ from a probability distribution μ in $\mathbb{R}^D$ representing the training data, an algorithm "learns" a dictionary $\widehat{\Phi}$ which provides sparse representations for the observations sampled from μ. This problem and its optimal algorithmic solutions are far from being well-understood, at least compared to the understanding that we have for classical dictionaries such as Fourier, wavelets, curvelets, and shearlets. These dictionaries arise in computational harmonic analysis approaches to image processing, and Donoho (1999) (for example) provides rigorous, optimal approximation results for simple classes of images. The work of Gribonval et al. (2013) presents rather general bounds for the complexity of learning the dictionaries (see also Maurer and Pontil, 2010b; Vainsencher et al., 2011, and references therein). The algorithms used in dictionary learning are often computationally demanding, and many of them are based on high-dimensional non-convex optimization. The emphasis of existing work is often made on the generality of the approach, where minimal assumptions are made on geometry of the distribution from which the sample is generated. This "pessimistic" approach incurs bounds dependent upon the ambient dimension D in general (even in the standard case of data lying on one hyperplane).

A different type of geometric assumption on the data gives rise to manifold learning, where the observations aggregate on a *suitably regular manifold* $\mathcal{M}$ of dimension d isometrically embedded in $\mathbb{R}^D$ (notable works include Belkin and Niyogi, 2003; Coifman et al., 2005a,b; Coifman and Maggioni, 2006; Donoho and Grimes, 2003, 2002; Genovese et al., 2012b; Jones et al., 2008, 2010; Little et al., 2009; Little, 2012; Roweis and Saul, 2000; Tenenbaum et al., 2000; Zhang and Zha, 2002, among others). This setting has been recognized as useful in a variety of applications (e.g. Causevic et al., 2006; Coifman et al., 2006; Rahman et al., 2005)), influencing work in the applied mathematics and machine learning communities during the past several years. It has also been recognized that in many cases the data does not naturally aggregate on a smooth manifold (as in Little et al., 2009; Little, 2012; Wakin et al., 2005), with examples arising in imaging that contradict the smoothness conditions. While this phenomenon is not as widely recognized as it probably should be, we believe that it is crucial to develop methods (both for dictionary and manifold learning) that are robust not only to noise, but also to modeling error. Such concerns motivated

the work on intrinsic dimension estimation of noisy data sets (see Little et al., 2009; Little, 2012), where smoothness of the underlying distribution of the data is not assumed, but only certain natural conditions (possibly varying with the scale of the data) are imposed. The central idea of the aforementioned works is to perform the multiscale singular value decomposition (SVD) of the data, an approach inspired by the works of David and Semmes (1993) and Jones (1990) in classical geometric measure theory. These techniques were further extended in several directions in the papers by Chen and Maggioni (2011); Chen et al. (2011a,b), while Allard et al. (2012); Chen and M.Maggioni (2010) built upon this work to construct multiscale dictionaries for the data based on the idea of Geometric Multi-Resolution Analysis (GMRA).

Until these recent works introduced the GMRA construction, connections between dictionary learning and manifold learning had not garnered much attention in the literature. These papers showed that, for intrinsically low-dimensional data, one may perform dictionary learning very efficiently by exploiting the underlying geometry, thereby illuminating the relationship between manifold learning and dictionary learning. In these papers, it was demonstrated that, in the infinite sample limit and under a manifold model assumption for the distribution of the data (with mild regularity conditions for the manifold), the GMRA algorithm efficiently learns a dictionary in which the data admits sparse representations. More interestingly, the examples in that paper show that the GMRA construction succeeds on real-world data sets which do not admit a structure consistent with the smooth manifold modeling assumption, suggesting that the GMRA construction exhibits robustness to modeling error. This desirable behavior follows naturally from design decisions; GMRA combines two elements that add stability: a multiscale decomposition and localized SVD.

In this paper, we analyze the finite sample behavior of (a slightly modified version of) that construction, and prove strong finite-sample guarantees for its behavior under general conditions on the geometry of a probability distribution generating the data. In particular, we show that these conditions are satisfied when the probability distribution is concentrated "near" a manifold, which robustly accounts for noise and modeling errors. In contrast to the pessimistic bounds mentioned above, the bounds that we prove only depend on the "intrinsic dimension" of the data. It should be noted that our method of proof produces non-asymptotic bounds, and requires several explicit geometric arguments not previously available in the literature (at least to the best of our knowledge). Some of our geometric bounds could be of independent interest to the manifold learning community.

The GMRA construction is therefore proven to simultaneously learn manifolds (in sense that it outputs a suitably close approximation to points on a manifold) and dictionaries in which the data is represented sparsely. Moreover, the construction is guaranteed to be robust with respect to noise and to perturbations of the manifold model. The GMRA construction is fast, linear in the size of the data matrix, inherently online, does not require nonlinear optimization, and is not iterative. Finally, our results may be combined with the GMRA compressed sensing techniques and algorithms presented in Iwen and Maggioni (2013), yielding both a method to learn a dictionary in a stable way on a finite set of training data, and a way of performing compressive sensing and reconstruction (with guarantees) from a small number of (suitable) linear projections (again without the need for expensive convex optimization).

This paper is organized as follows: Section 2 introduces the main definitions and notation employed throughout the paper. Section 3 explains the main contributions, formally states the results and provides comparison with existing literature. Finally, Sections 4 and 5 are devoted to the proofs of our main results, Theorem 2 and Theorem 6.

2. Geometric Multi-Resolution Analysis (GMRA)

This section describes the mathematical framework and the main objects studied in the paper. Our goal will be to prove the following claims which we explain informally at this point. In the statement below, "$\gtrsim$" and "$\lesssim$" denote inequalities up to multiplicative constants and logarithmic factors. The results will be made formal in the course of exposition.

Statement of results. *Let $\sigma \geq 0$ be a fixed small constant, and let $\varepsilon \gtrsim \sigma$ be given. Suppose that $n \gtrsim \varepsilon^{-(1+d/2)}$, and let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ be an i.i.d. sample from Π, a probability distribution with density supported in a tube of radius σ around a smooth closed d-dimensional manifold $\mathcal{M} \hookrightarrow \mathbb{R}^D$. There exists an algorithm that, given $\mathcal{X}_n$, outputs the following objects:*

- *a dictionary $\widehat{\Phi}_\varepsilon = \{\widehat{\varphi}_i\}_{i \in \mathcal{J}_\varepsilon} \subset \mathbb{R}^D$;*
- *a nonlinear "encoding" operator $\widehat{\mathcal{D}}_\varepsilon : \mathbb{R}^D \to \mathbb{R}^{\mathcal{J}_\varepsilon}$ which takes $x \in \mathbb{R}^D$ and returns the coefficients of its approximation by the elements of $\widehat{\Phi}_\varepsilon$;*
- *a "decoding" operator $\widehat{\mathcal{D}}_\varepsilon^{-1} : \mathbb{R}^{\mathcal{J}_\varepsilon} \to \mathbb{R}^D$ which maps a sequence of coefficients to an element of $\mathbb{R}^D$.*

Moreover, the following properties hold with high probability:

i. $\mathrm{Card}(\mathcal{J}_\varepsilon) \lesssim \varepsilon^{-d/2}$;

ii. the image of $\widehat{\mathcal{D}}_\varepsilon$ is contained in the set $S_{d+1} \subset \mathbb{R}^{\mathcal{J}_\varepsilon}$ of all $(d+1)$ - sparse vectors (i.e., vectors with at most $d+1$ nonzero coordinates);

iii. the reconstruction error satisfies

$$\sup_{x \in \mathrm{support}(\Pi)} \|x - \widehat{\mathcal{D}}_\varepsilon^{-1} \widehat{\mathcal{D}}_\varepsilon(x)\| \lesssim \varepsilon;$$

iv. the time complexity for computing

- *$\widehat{\Phi}_\varepsilon$ is $O(C^d(D+d^2)\varepsilon^{-1-\frac{2}{d}}\log(1/\varepsilon))$, where C is a universal constant;*
- *$\widehat{\mathcal{D}}_\varepsilon(x)$ is $O(d(D + d\log(1/\varepsilon)))$, and for $\widehat{\mathcal{D}}_\varepsilon^{-1}(x)$ is $O(d(D+\log(1/\varepsilon)))$.*

If a new observation X_{n+1} from Π becomes available, $\widehat{\Phi}_\varepsilon$ may be updated in time $O(C^d(D+d^2)\log(1/\varepsilon))$.

In other words, we can construct a data-dependent dictionary $\widehat{\Phi}_\varepsilon$ of cardinality $O(\varepsilon^{-d/2})$ by looking at $n \asymp \varepsilon^{-1-\frac{2}{d}}$ data points drawn from Π, such that $\widehat{\Phi}_\varepsilon$ provides both $(d+1)$-sparse approximations to data and has expected "reconstruction error" of order ε (with high probability). Moreover, the algorithm producing this dictionary is fast and can be quickly

updated if new points become available. We want to emphasize that the complexity of our construction only depends on the desired accuracy ε, and is independent of the total number of samples (more precisely, it is enough to use only the first $\simeq \varepsilon^{-(1+d/2)}$ data points). Many existing techniques in dictionary learning cannot guarantee a requested accuracy, or a given sparsity, and a certain computational cost as a function of the two. Our results above completely characterize the tradeoffs between desired precision, dictionary size, sparsity, and computational complexity for our dictionary learning procedure.

We also remark that a suitable version of compressed sensing applies to the dictionary representations used in the theorem: we refer the reader to the works by Chen et al. (2012); Iwen and Maggioni (2013).

2.1 Notation

For $v \in \mathbb{R}^D$, $\|v\|$ denotes the standard Euclidean norm in $\mathbb{R}^D$. $B_d(0,r)$ is the Euclidean ball in $\mathbb{R}^d$ of radius r centered at the origin, and we let $B(0,r) := B_D(0,r)$. Proj_V stands for the orthogonal projection onto a linear subspace $V \subseteq \mathbb{R}^D$, $\dim(V)$ for its dimension and $V^\perp$ for its orthogonal complement. For $x \in \mathbb{R}^D$, let Proj_{x+V} be the affine projection onto the affine subspace $x+V$ defined by $\mathrm{Proj}_{x+V}(y) = x + \mathrm{Proj}_V(y-x)$, for $y \in \mathbb{R}^D$.

Given a matrix $A \in \mathbb{R}^{k\times l}$, we write $A = [a_1|\cdots|a_l]$, where a_i stands for the ith column of A. The operator norm is denoted by $\|A\|$, the Frobenius norm by $\|A\|_F$ and the matrix transpose by A^T. If $k = l$, $\mathrm{tr}\,(A)$ denotes the trace. For $v \in \mathbb{R}^k$, let $\mathrm{diag}(v)$ be the $k \times k$ diagonal matrix with $(\mathrm{diag}(v))_{ii} = v_i,\ i = 1,\dots,k$. Finally, we use $\mathrm{span}\{a_i\}_{i=1}^l$ to denote the linear span of the columns of A.

Given a C^2 function $f : \mathbb{R}^l \to \mathbb{R}^k$, let f_i denote the ith coordinate of the function f for $i = 1,\dots k$, $Df(v)$ the Jacobian of f at $v \in \mathbb{R}^l$, and $D^2 f_i(v)$ the Hessian of the ith coordinate at v.

We shall use $d\mathrm{Vol}$ to denote Lebesgue measure on $\mathbb{R}^D$, and if $U \subset \mathbb{R}^D$ is Lebesgue measurable, $\mathrm{Vol}(U)$ stands for the Lebesgue measure of U. We will use $\mathrm{Vol}_{\mathcal{M}}$ to denote the volume measure on a d-manifold $\mathcal{M}$ in $\mathbb{R}^D$ (note that this coincides with the d-dimensional Hausdorff measure for the subset $\mathcal{M}$ of $\mathbb{R}^D$), and $d_{\mathcal{M}}(x,y)$ to denote the geodesic distance between two points $x, y \in \mathcal{M}$. For a probability measure Π on $\mathbb{R}^D$,

$$\mathrm{supp}(\Pi) := \bigcap_{\mathcal{C} \text{ closed}, \Pi(C)=1} \mathcal{C}$$

stands for its support. Finally, for $x, y \in \mathbb{R}$, $x \vee y := \max(x,y)$.

2.2 Definition of the geometric multi-resolution analysis (GMRA)

We assume that the data are identically, independently distributed samples from a Borel probability measure Π on $\mathbb{R}^D$. Let $1 \le d \le D$ be an integer. A GMRA with respect to the probability measure Π consists of a collection of (nonlinear) operators $\{P_j : \mathbb{R}^D \to \mathbb{R}^D\}_{j\ge 0}$. For each "resolution level" $j \ge 0$, P_j is uniquely defined by a collection of pairs of subsets and affine projections, $\{(C_{j,k}, P_{j,k})\}_{k=1}^{N(j)}$, where the subsets $\{C_{j,k}\}_{k=1}^{N(j)}$ form a measurable partition of $\mathbb{R}^D$ (that is, members of $\{C_{j,k}\}_{k=1}^{N(j)}$ are pairwise disjoint and the union of all

members is $\mathbb{R}^D$). P_j is constructed by piecing together local affine projections. Namely, let

$$P_{j,k}(x) := c_{j,k} + \text{Proj}_{V_{j,k}}(x - c_{j,k}),$$

where $c_{j,k} \in \mathbb{R}^D$ and $V_{j,k}$ are defined as follows. Let $\mathbb{E}_{j,k}$ stand for the expectation with respect to the conditional distribution $d\Pi_{j,k}(x) = d\Pi(x|x \in C_{j,k})$. Then

$$c_{j,k} = \mathbb{E}_{j,k} x, \tag{1}$$

$$V_{j,k} = \underset{\dim(V)=d}{\text{argmin}} \; \mathbb{E}_{j,k} \left\| x - c_{j,k} - \text{Proj}_V(x - c_{j,k}) \right\|^2, \tag{2}$$

where the minimum is taken over all linear spaces V of dimension d. In other words, $c_{j,k}$ is the conditional mean and $V_{j,k}$ is the subspace spanned by eigenvectors corresponding to d largest eigenvalues of the conditional covariance matrix

$$\Sigma_{j,k} = \mathbb{E}_{j,k}(x - c_{j,k})(x - c_{j,k})^T. \tag{3}$$

Note that we have implicitly assumed that such a subspace $V_{j,k}$ is unique, which will always be the case throughout this paper. Given such a $\{(C_{j,k}, P_{j,k})\}_{k=1}^{N(j)}$, we define

$$P_j(x) := \sum_{k=1}^{N(j)} I\{x \in C_{j,k}\} P_{j,k}(x)$$

where $I\{x \in C_{j,k}\}$ is the indicator function of the set $C_{j,k}$.

It was shown in the paper by Allard et al. (2012) that if Π is supported on a smooth, closed d-dimensional submanifold $\mathcal{M} \hookrightarrow \mathbb{R}^D$, and if the partitions $\{C_{j,k}\}_{j=1}^{N(j)}$ satisfy some regularity conditions for each j, then, for any $x \in \mathcal{M}$, $\|x - P_j(x)\| \leq C(\mathcal{M}) 2^{-2j}$ for all $j \geq j_0(\mathcal{M})$. This means that the operators P_j provide an efficient "compression scheme" $x \mapsto P_j(x)$ for $x \in \mathcal{M}$, in the sense that every x can be well-approximated by a linear combination of at most $d+1$ vectors from the dictionary $\Phi_{2^{-2j}}$ formed by $\{c_{j,k}\}_{k=1}^{N(j)}$ and the union of the bases of $V_{j,k}$, $k = 1 \ldots N(j)$. Furthermore, operators efficiently encoding the "difference" between P_j and P_{j+1} were constructed, leading to a multiscale compressible representation of $\mathcal{M}$.

In practice, Π is unknown and we only have access to the *training data* $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, which are assumed to be i.i.d. with distribution Π. In this case, operators P_j are replaced by their estimators

$$\widehat{P}_j(x) := \sum_{k=1}^{N(j)} I\{x \in C_{j,k}\} \widehat{P}_{j,k}(x)$$

where $\{C_{j,k}\}_{k=1}^{N(j)}$ is a suitable partition of $\mathbb{R}^D$ *obtained from the data*,

$$\widehat{P}_{j,k}(x) := \widehat{c}_{j,k} + \text{Proj}_{\widehat{V}_{j,k}}(x - \widehat{c}_{j,k}), \tag{4}$$

$$\widehat{c}_{j,k} := \frac{1}{|\mathcal{X}_{j,k}|} \sum_{x \in \mathcal{X}_{j,k}} x,$$

$$\widehat{V}_{j,k} := \underset{\dim(V)=d}{\text{argmin}} \frac{1}{|\mathcal{X}_{j,k}|} \sum_{x \in \mathcal{X}_{j,k}} \left\| x - \widehat{c}_{j,k} - \text{Proj}_V(x - \widehat{c}_{j,k}) \right\|^2,$$

$\mathcal{X}_{j,k} = C_{j,k} \cap \mathcal{X}_n$, and $|\mathcal{X}_{j,k}|$ denotes the number of elements in $\mathcal{X}_{j,k}$. We shall call these $\widehat{P}_j$ the *empirical GMRA*.

Moreover, the dictionary $\widehat{\Phi}_{2^{-2j}}$ is formed by $\{\hat{c}_{j,k}\}_{k=1}^{N(j)}$ and the union of bases of $\hat{V}_{j,k}$, $k = 1 \dots N(j)$. The "encoding" and "decoding" operators $\widehat{\mathcal{D}}_{2^{-2j}}$ and $\widehat{\mathcal{D}}^{-1}_{2^{-2j}}$ mentioned above are now defined in the obvious way, so that $\widehat{\mathcal{D}}^{-1}_{2^{-2j}}\widehat{\mathcal{D}}_{2^{-2j}}(x) = \widehat{P}_{j,k}(x)$ for any $x \in C_{j,k}$.

We remark that the "intrinsic dimension" d is assumed to be known throughout this paper. In practice, it can estimated within the GMRA construction using the "multiscale SVD" ideas of Little et al. (2009); Little (2012). The estimation technique is based on inspecting (for a given point $x \in C_{j,k}$) the behavior of the singular values of the covariance matrix $\Sigma_{j,k}$ as j varies. For alternative methods, see Camastra and Vinciarelli (2001); Levina and Bickel (2004) and references therein.

3. Main results

Our main goal is to obtain the *non-asymptotic* probabilistic bounds on the performance of the empirical GMRA under certain structural assumptions on the underlying distribution of the data. In practice, the data rarely belongs precisely to a smooth low-dimensional submanifold. One way to relax this condition is to assume that it is "sufficiently close" to a nice set. Here we assume that the underlying distribution is supported in a thin tube around the manifold. We make no assumptions about the structure or distribution of the noise, instead trying to understand how the error of sparse approximation depends on the "thickness" of the tube, which quantifies stability and robustness properties of our algorithm. Another way to model this situation is to allow *additive noise*, whence the observations are assumed to be of the form $X = Y + \xi$, where Y belongs to a submanifold of $\mathbb{R}^D$, ξ is independent of Y, and the distribution of ξ is known. This leads to a singular deconvolution problem (see Genovese et al., 2012c; Koltchinskii, 2000).

We will measure performance of the empirical GMRA by the $L_2(\Pi)$-error

$$\mathbb{E}\left\|X - \widehat{P}_j(X)\right\|^2 := \int\limits_{\mathrm{supp}(\Pi)} \left\|x - \widehat{P}_j(x)\right\|^2 d\Pi(x) \tag{5}$$

or by the $\|\cdot\|_{\infty,\Pi}$-error defined as

$$\left\|\mathrm{Id} - \widehat{P}_j\right\|_{\infty,\Pi} := \sup_{x \in \mathrm{supp}(\Pi)} \left\|x - \widehat{P}_j(x)\right\|, \tag{6}$$

where $\widehat{P}_j$ is defined by (4). As we mentioned before, our GMRA construction is entirely data-dependent: it takes the point cloud of cardinality n as an input and for every $j \in \mathbb{Z}_+$ returns the partition $\{C_{j,k}\}_{k=1}^{N(j)}$ and associated affine projectors $\widehat{P}_{j,k}$.

The presentation is structured as follows: we start from the natural decomposition

$$\left\|x - \widehat{P}_j(x)\right\| \le \underbrace{\|x - P_j(x)\|}_{\text{approximation error}} + \underbrace{\left\|P_j(x) - \widehat{P}_j(x)\right\|}_{\text{random error}}$$

and state the general conditions on the underlying distribution and partition scheme that suffice to guarantee that

1. the distribution-dependent operators P_j yield good approximation, as measured by $\mathbb{E}\,\|x - P_j(x)\|^2$;

2. the empirical version $\widehat{P}_j$ is with high probability close to P_j, so that $\mathbb{E}\left\|\widehat{P}_j(x) - P_j(x)\right\|^2$ is small.

This leads to our first result, Theorem 2, where the error $\mathbb{E}\left\|x - \widehat{P}_j(x)\right\|^2$ of the empirical GMRA is bounded with high probability.

After developing this general result, we then consider the special but important case where the distribution Π generating the data is supported in thin tube around a smooth submanifold, and for a (concrete, efficiently computable, online) partition scheme we show that the conditions of Theorem 2 are satisfied. This is summarized in the statement of Theorem 6, that may be interpreted as proving finite-sample bounds for our GMRA-based dictionary learning scheme for high-dimensional data that suitably concentrates around a manifold. It is important to note that most of the constants in our results are explicit. The only geometric parameters involved in the bounds are the dimension d of the manifold (but not the ambient dimension D), its *reach* (see τ in (9)) and the "tube thickness" σ.

Among the existing literature, the papers Allard et al. (2012); Chen et al. (2012) introduced the idea of using multiscale geometric decomposition of data to estimate the distribution of points sampled in high-dimensions. However in the first paper no finite sample analysis was performed, and in the second the connection with geometric properties of the distribution of the data is not made explicit, and the conditions are expressed in terms of certain approximation spaces within the space of probability distributions in $\mathbb{R}^D$, with Wasserstein metrics used to measure distances and approximation errors.

The recent paper by Canas et al. (2012) is close in scope to our work: its authors present probabilistic guarantees for approximating a manifold with a global solution of the so-called k-flats (Bradley and Mangasarian, 2000) problem in the case of distributions supported on manifolds. It is important to note, however, that our estimator is explicitly computable, while exact global solution of k-flats is usually unavailable and certain approximations are used in practice, and convergence to a global minimum is conditioned on suitable unknown initializations. We also seamlessly tackle the case of noise and model error, which is beyond what was studied previously. We consider this development extremely relevant in applications, both because real data is corrupted by noise and the assumption that data lies exactly on a smooth manifold is often unrealistic. A more detailed comparison of theoretical guarantees for k-flats and for our approach is given after we state the main results in Subsection 3.2 below.

Another body of literature connected to this work studies the complexity of dictionary learning. For example, Gribonval et al. (2013) present rather general bounds for the complexity of learning the dictionaries (those results build on and generalize the works of Maurer and Pontil (2010b); Vainsencher et al. (2011), among several others). The emphasis of that work is on the generality of the approach, at the expense of obtaining bounds that are rather pessimistic in general (even in the standard case of data lying on one hyperplane) and depend on the ambient dimension D of the problem, while the bounds we present only depend on the intrinsic dimension of the data.

In the course of the proof, we obtain several results that might be of independent interest. In particular, Lemma 16 gives upper and lower bounds for the volume of the tube around a manifold in terms of the reach (7) and tube thickness. While the exact tubular volumes are given by Weyl's tube formula (see Gray, 2004), our bound are exceedingly easy to state in terms of simple global geometric parameters.

For the details on numerical implementation of GMRA and its modifications, see the works by Allard et al. (2012); Chen and M.Maggioni (2010).

3.1 Finite sample bounds for empirical GMRA

In this section, we shall present the finite sample bounds for the empirical GMRA described above. For a fixed resolution level j, we first state sufficient conditions on the distribution Π and the partition $\{C_{j,k}\}_{k=1}^{N(j)}$ for which these $L_2(\Pi)$-error bounds hold (see Theorem 2 below).

Suppose that for all integers $j_{\min} \leq j \leq j_{\max}$ the following is true:

(A1) There exists an integer $1 \leq d \leq D$ and a positive constant $\theta_1 = \theta_1(\Pi)$ such that for all $k = 1, \dots, N(j)$,

$$\Pi(C_{j,k}) \geq \theta_1 2^{-jd}.$$

(A2) There is a positive constant $\theta_2 = \theta_2(\Pi)$ such that for all $k = 1, \dots, N(j)$, if X is drawn from $\Pi_{j,k}$ then, Π - almost surely,

$$\|X - c_{j,k}\| \leq \theta_2 2^{-j}.$$

(A3) Let $\lambda_1^{j,k}, \dots, \lambda_D^{j,k}$ denote the eigenvalues of the covariance matrix $\Sigma_{j,k}$ (defined in 3) arranged in the non-increasing order. Then there exist $\sigma = \sigma(\Pi) \geq 0$, $\theta_3 = \theta_3(\Pi)$, $\theta_4 = \theta_4(\Pi) > 0$, and some $\alpha \in (0, 1]$ such that for all $k = 1 \dots N(j)$,

$$\lambda_d^{j,k} \geq \theta_3 \frac{2^{-2j}}{d} \quad \text{and} \quad \sum_{l=d+1}^{D} \lambda_l^{j,k} \leq \theta_4(\sigma^2 + 2^{-2(1+\alpha)j}) \leq \frac{1}{2}\lambda_d^{j,k}.$$

If in addition

(A4) There exists $\theta_5 = \theta_5(\Pi)$ such that

$$\|\mathrm{Id} - P_j\|_{\infty,\Pi} \leq \theta_5 \left(\sigma + 2^{-(1+\alpha)j}\right),$$

then the bounds are also guaranteed to hold for the $\|\cdot\|_{\infty,\Pi}$-error (6).

Remark 1

i. Assumption (A1) *entails that the distribution assigns a reasonable amount of probability to each partition element, assumption* (A2) *ensures that samples from partition elements are always within a ball around the centroid, and assumption* (A3) *controls the effective dimensionality of the samples within each partition element. Assumption* (A4) *just assumes a bound on the error for the theoretical GMRA reconstruction.*

ii. *Note that the constants θ_i, $i = 1 \dots 4$, are independent of the resolution level j.*

iii. *It is easy to see that Assumption (A3) implies a bound on the "local approximation error": since P_j acts on $C_{j,k}$ as an affine projection on the first d "principal components", we have*

$$\begin{aligned}\mathbb{E}_{j,k}\|x - P_j(x)\|^2 &= \operatorname{tr}\left[\mathbb{E}_{j,k}\left(x - c_{j,k} - \operatorname{Proj}_{V_{j,k}}(x)\right)\left(x - c_{j,k} - \operatorname{Proj}_{V_{j,k}}(x)\right)^T\right] \\ &= \sum_{l=d+1}^{D} \lambda_l^{j,k} \le \theta_4(\sigma^2 + 2^{-2(1+\alpha)j}).\end{aligned}$$

iv. *The parameter σ is introduced to cover "noisy" models, including the situations when Π is supported in a thin tube of width σ around a low-dimensional manifold $\mathcal{M}$. Whenever Π is supported on a smooth d-dimensional manifold, σ can be taken to be 0.*

v. *The stipulation*

$$\theta_4(\sigma^2 + 2^{-2(1+\alpha)j}) \le \frac{1}{2}\lambda_d^{j,k}$$

guarantees that the spectral gap $\lambda_d^{j,k} - \lambda_{d+1}^{j,k}$ is sufficiently large.

We are in position to state our main result.

Theorem 2 *Suppose that **(A1)-(A3)** are satisfied, let $X, X_1, \dots, X_n$ be an i.i.d. sample from Π, and set $\bar{d} := 4d^2\theta_2^4/\theta_3^2$. Then for any $j_{\min} \le j \le j_{\max}$ and any $t \ge 1$ such that $t + \log(\bar{d} \vee 8) \le \frac{1}{2}\theta_1 n 2^{-jd}$,*

$$\mathbb{E}\|X - \widehat{P}_j(X)\|^2 \le 2\theta_4\left(\sigma^2 + 2^{-2j(1+\alpha)}\right) + c_1 2^{-2j}\frac{(t + \log(\bar{d} \vee 8))d^2}{n2^{-jd}},$$

*and if in addition **(A4)** is satisfied,*

$$\left\|\mathrm{Id} - \widehat{P}_j\right\|_{\infty,\Pi} \le \theta_5\left(\sigma + 2^{-(1+\alpha)j}\right) + \sqrt{\frac{c_1}{2}2^{-2j}\frac{(t + \log(\bar{d} \vee 8))d^2}{n2^{-jd}}}$$

with probability $\ge 1 - \frac{2^{jd+1}}{\theta_1}\left(e^{-t} + e^{-\frac{\theta_1}{16}n2^{-jd}}\right)$, where $c_1 = 2\left(12\sqrt{2}\frac{\theta_2^3}{\theta_3\sqrt{\theta_1}} + 4\sqrt{2}\frac{\theta_2}{d\sqrt{\theta_1}}\right)^2$.

3.2 Distributions concentrated near smooth manifolds

Of course, the statement of Theorem 2 has little value unless assumptions **(A1)-(A4)** can be verified for a rich class of underlying distributions. We now introduce an important class of models and an algorithm to construct suitable partitions $\{C_{j,k}\}$ which together satisfy these assumptions. Let $\mathcal{M}$ be a smooth (or at least C^2, so changes of coordinate charts admit continuous second-order derivatives), closed d-dimensional submanifold of $\mathbb{R}^D$. We

recall the definition of the *reach* (see Federer, 1959), an important global characteristic of $\mathcal{M}$. Let

$$D(\mathcal{M}) = \{y \in \mathbb{R}^D : \exists! x \in \mathcal{M} \text{ s.t. } \|x - y\| = \inf_{z \in \mathcal{M}} \|z - y\|\}, \tag{7}$$

$$\mathcal{M}_r = \{y \in \mathbb{R}^D : \inf_{x \in \mathcal{M}} \|x - y\|\} < r\}. \tag{8}$$

Then

$$\text{reach}(\mathcal{M}) := \sup\{r \geq 0 : \ \mathcal{M}_r \subseteq D(\mathcal{M})\}, \tag{9}$$

and we shall always use τ to denote the reach of the manifold $\mathcal{M}$.

Definition 3 *Assume that $0 \leq \sigma < \tau$. We shall say that the distribution Π satisfies the* **(τ, σ)-model assumption** *if there exists a bounded smooth (or at least C^2) submanifold $\mathcal{M} \hookrightarrow \mathbb{R}^D$ with reach τ such that $\text{supp}(\Pi) = \mathcal{M}_\sigma$, Π is absolutely continuous with respect to $U_{\mathcal{M}_\sigma}$ - the uniform distribution on $\mathcal{M}_\sigma$ - and the Radon-Nikodym derivative $\frac{d\Pi}{dU_{\mathcal{M}_\sigma}}$ satisfies*

$$0 < \phi_1 \leq \frac{d\Pi}{dU_{\mathcal{M}_\sigma}} \leq \phi_2 < \infty \qquad U_{\mathcal{M}_\sigma} \text{ almost surely}. \tag{10}$$

Our partitioning scheme is based on the data structure known as the *cover tree* introduced by Beygelzimer et al. (2006) (see also Ciaccia et al., 1997; Karger and Ruhl, 2002; Yianilos, 1993). We briefly recall its definition and basic properties. Given a set of n distinct points $S_n = \{x_1, \ldots, x_n\}$ in some metric space (S, ρ), the cover tree T on S_n satisfies the following: let $T_j \subset S_n$, $j = 0, 1, 2, \ldots$ be the set of nodes of T at level j. Then

1. $T_j \subset T_{j+1}$;
2. for all $y \in T_{j+1}$, there exists $z \in T_j$ such that $\rho(y, z) < 2^{-j}$;
3. for all $y, z \in T_j$, $\rho(y, z) > 2^{-j}$.

Remark 4 *Note that these properties imply the following: for any $y \in S_n$, there exists $z \in T_j$ such that $\rho(y, z) < 2^{-j+1}$.*

Theorem 3 in (Beygelzimer et al., 2006) shows that the cover tree always exists; for more details, see the aforementioned paper.

We will construct a cover tree for the collection $X_1, \ldots, X_n$ of i.i.d. samples from the distribution Π with respect to the Euclidean distance $\rho(x, y) := \|x - y\|$. Assume that $T_j := T_j(X_1, \ldots, X_n) = \{a_{j,k}\}_{k=1}^{N(j)}$. Define the indexing map

$$k(x) := \operatorname*{argmin}_{1 \leq k \leq N(j)} \|x - a_{j,k}\|$$

(ties are broken by choosing the smallest value of k), and partition $\mathbb{R}^D$ into the Voronoi regions

$$C_{j,k} = \{x \in \mathbb{R}^D : k_j(x) = k\}. \tag{11}$$

Let $\varepsilon(n,t)$ be the smallest $\varepsilon > 0$ which satisfies

$$n \geq \frac{1}{\phi_1}\left(\frac{\tau+\sigma}{\tau-\sigma}\right)^d \beta_1 \left(\log \beta_2 + t\right), \tag{12}$$

where $\beta_1 = \frac{\operatorname{Vol}_{\mathcal{M}}(\mathcal{M})}{\cos^d(\delta_1)\operatorname{Vol}(B_d(0,\varepsilon/4))}$, $\beta_2 = \frac{\operatorname{Vol}_{\mathcal{M}}(\mathcal{M})}{\cos^d(\delta_2)\operatorname{Vol}(B_d(0,\varepsilon/8))}$, $\delta_1 = \arcsin(\varepsilon/8\tau)$, and $\delta_2 = \arcsin(\varepsilon/16\tau)$.

Remark 5 *For large enough n, this requirement translates into $n \geq C(\mathcal{M}, d, \phi_1)\left(\frac{1}{\varepsilon}\right)^d \left(\log \frac{1}{\varepsilon} + t\right)$ for some constant $C(\mathcal{M}, d, \phi_1)$.*

We are ready to state the main result of this section.

Theorem 6 *Suppose that Π satisfies the (τ,σ)-model assumption. Let $X_1,\dots,X_n$ be an i.i.d. sample from Π, construct a cover tree T from $\{X_i\}_{i=1}^n$, and define $C_{j,k}$ as in (11). Assume that $\varepsilon(n,t) < \sigma$. Then, for all $j \in \mathbb{Z}_+$ such that $2^{-j} > 8\sigma$ and $3\cdot 2^{-j} + \sigma < \tau/8$, partition $\{C_{j,k}\}_{k=1}^{N(j)}$ and Π satisfy **(A1)**, **(A2)**, **(A3)**, and **(A4)** with probability $\geq 1 - e^{-t}$ for*

$$\begin{aligned}
\theta_1 &= \frac{\phi_1 \operatorname{Vol}(B_d(0,1))}{2^{4d}\operatorname{Vol}_{\mathcal{M}}(\mathcal{M})}\left(\frac{\tau-\sigma}{\tau+\sigma}\right)^d,\\
\theta_2 &= 12,\\
\theta_3 &= \frac{\phi_1/\phi_2}{2^{4d+8}\left(1+\frac{\sigma}{\tau}\right)^d},\\
\theta_4 &= 2 \vee \frac{2^3 3^4}{\tau^2},\\
\theta_5 &= \left(2 \vee \frac{2^2 3^2}{\tau}\right)\left(1 + 3\cdot 2^5\sqrt{2d}\left(1+\frac{\sigma}{\tau}\right)^{d/2}\left(\frac{1+\left(\frac{25}{71}\right)^2}{1-\frac{1}{9\cdot 2^{12}}}\right)^{d/4}\right),\\
\alpha &= 1.
\end{aligned}$$

One may combine the results of Theorem 6 and Theorem 2 as follows: given an i.i.d. sample $X_1,\dots,X_n$ from Π, use the first $\lceil \frac{n}{2} \rceil$ points $\{X_1,\dots,X_{\lceil \frac{n}{2}\rceil}\}$ to obtain the partition $\{C_{j,k}\}_{k=1}^{N(j)}$, while the remaining $\{X_{\lceil \frac{n}{2}\rceil+1},\dots,X_n\}$ are used to construct the operator $\hat{P}_j$ (see (4)). This makes our GMRA construction entirely (cover tree, partitions, affine linear projections) data-dependent.

In the special case when $\sigma = 0$, the bounds resulting from Theorem 2 can be "optimized" over j to get the following statement (we present only the bounds for the $L_2(\Pi)$ error, but the results $\|\cdot\|_{\infty,\Pi}$ are similar).

Corollary 7 *Assume that conditions of Theorem 6 hold with $\sigma = 0$, and that n is sufficiently large. Then for all $A \geq 1$ such that $A\log n \leq c_4 n$, the following holds:*

(a) *if $d \in \{1,2\}$,*

$$\inf_{j\in\mathbb{Z}:2^{-j}<\tau/24} \mathbb{E}\|x - \widehat{P}_j(x)\|^2 \leq C_1\left(\frac{\log n}{n}\right)^{\frac{2}{d}};$$

(b) *if* $d \geq 3$,

$$\inf_{j \in \mathbb{Z}: 2^{-j} < \tau/24} \mathbb{E}\|x - \widehat{P}_j(x)\|^2 \leq C_2 \left(\frac{\log n}{n}\right)^{\frac{4}{d+2}} \tag{13}$$

with probability $\geq 1 - c_3 n^{-A}$, *where* C_1 *and* C_2 *depend only on* $A, \tau, d, \phi_1/\phi_2, \mathrm{Vol}_{\mathcal{M}}(\mathcal{M})$ *and* c_3, c_4 *depend only on* $\tau, d, \phi_1/\phi_2, \mathrm{Vol}_{\mathcal{M}}(\mathcal{M})$.

Proof In case (a), it is enough to set $t := (A+1)\log n$, $2^{-j} := \left(\frac{16t}{\theta_1 n}\right)^{1/d}$, and apply Theorem 2. For case (b), set $t := (A+1)\log n$ and $2^{-j} := \left(\frac{A \log n}{n}\right)^{\frac{1}{d+2}}$. ■

Finally, we note that the claims *ii.* and *iii.* stated in the beginning of Section 2 easily follow from our general results (it is enough to choose n such that $\varepsilon \simeq n^{-\frac{2}{d+2}}$ and $2^{-j} = \sqrt{\varepsilon}$). Claim *i.* follows from assumption **(A1)** and Theorem 6. Computational complexity bounds *iv.* follow from the associated computational cost estimates for the cover trees algorithm and the randomized singular value decomposition, and are discussed in detail in Sections 3 and 8 of (Allard et al., 2012).

3.3 Connections to the previous work

It is useful to compare our rates with results of Theorem 4 in (Canas et al., 2012). In particular, this theorem implies that, given a sample of size n from the Borel probability measure Π on the smooth d-dimensional manifold $\mathcal{M}$, the $L_2(\Pi)$-error of approximation of $\mathcal{M}$ by $k_n = C_1(\mathcal{M}, \Pi) n^{d/(2(d+4))}$ affine subspaces is bounded by $C_2(\mathcal{M}, \Pi) n^{-2/(d+4)}$. Here, the dependence of k_n on n is "optimal" in a sense that it minimizes the upper bound for the risk obtained in (Canas et al., 2012). If we set $\sigma = 0$ in our results, then it easily follows from Theorems 6 and 2 that the $L_2(\Pi)$-error achieved by our GMRA construction for $2^j \simeq n^{\frac{1}{2(d+4)}}$ (so that $N(j) \simeq k_n$ to make the results comparable) is of the same order $n^{-\frac{2}{d+4}}$. However, this choice of j is not optimal in this case - in particular, setting $2^{j_n} \simeq n^{\frac{1}{d+2}}$, we obtain as in (13) a $L_2(\Pi)$-error of order $n^{-\frac{2}{d+2}}$, which is a faster rate. Moreover, we also obtain results in the sup norm, and not only for mean square error. We should note that technically our results require the stronger condition (10) on the underlying measure Π, while theoretical guarantees in (Canas et al., 2012) are obtained assuming only the upper bound $\frac{d\Pi}{dU_{\mathcal{M}}} \leq \phi_2 < \infty$, where $U_{\mathcal{M}} := \frac{d\,\mathrm{Vol}_{\mathcal{M}}}{\mathrm{Vol}_{\mathcal{M}}(\mathcal{M})}$ is the uniform distribution over $\mathcal{M}$.

The rate (13) is the same (up to log-factors) as the minimax rate obtained for the problem considered in (Genovese et al., 2012a) of estimating a manifold from the samples corrupted with the additive noise that is "normal to the manifold". Our theorems are stated under more general conditions, however, we only prove *robustness-type* results and do not address the problem of *denoising*. At the same time, the estimator proposed in (Genovese et al., 2012a) is (unlike our method) not suitable for applications. The paper (Genovese et al., 2012b) considers (among other problems) the noiseless case of manifold estimation under Hausdorff loss, and obtains the minimax rate of order $n^{-\frac{2}{d}}$. Performed numerical simulation (see Section 6) suggest that our construction also appears to achieve this rate in the noiseless case. However, we are interested in the much more realistic scenario of noisy data.

4. Preliminaries

This section contains the remaining definitions and preliminary technical facts that will be used in the proofs of our main results.

Given a point y on the manifold $\mathcal{M}$, let $T_y\mathcal{M}$ be the associated tangent space, and let $T_y^\perp\mathcal{M}$ be the orthogonal complement of $T_y\mathcal{M}$ in $\mathbb{R}^D$. We define the projection from the tube $\mathcal{M}_\sigma$ (see (8)) onto the manifold $\text{Proj}_{\mathcal{M}} : \mathcal{M}_\sigma \to \mathcal{M}$ by

$$\text{Proj}_{\mathcal{M}}(x) = \operatorname*{argmin}_{y\in\mathcal{M}} \|x-y\|$$

and note that $\sigma < \tau$, together with (7), implies that $\text{Proj}_{\mathcal{M}}$ is well-defined on $\mathcal{M}_\sigma$, and

$$\text{Proj}_{\mathcal{M}}(y+\xi) = y$$

whenever $y \in \mathcal{M}$ and $\xi \in T_y^\perp\mathcal{M} \cap B(0,\sigma)$.

Next, we recall some facts about the volumes of parallelotopes that will prove useful in Section 5. For a matrix $A \in \mathbb{R}^{k\times l}$ with $l \le k$, we shall abuse our previous notation and let $\text{Vol}(A)$ also denote the volume of the parallelotope formed by the columns of A. Let A and B be $k \times l_1$ and $k \times l_2$ matrices respectively with $l_1 + l_2 \le k$, and note that

$$\text{Vol}([A\,|\,B]) \le \text{Vol}(A)\text{Vol}(B)$$

where $([A\,|\,B])$ denotes the concatenation of A and B into a $k \times (l_1+l_2)$ matrix. Moreover, if the columns of A and B are all mutually orthogonal, we clearly have that $\text{Vol}([A\,|\,B]) = \text{Vol}(A)\text{Vol}(B)$. Assuming that I is the $l_1 \times l_1$ identity matrix, we have the bound $\text{Vol}\begin{pmatrix} A \\ I \end{pmatrix} \ge 1$. The following proposition gives volume bounds for specific types of perturbations that we shall encounter.

Proposition 8 *Suppose $Y = [y_1|\cdots|y_d]$ is symmetric d by d matrix such that $\|Y\| \le q < 1$. Then*

$$\begin{aligned} \text{Vol}\begin{pmatrix} I+Y \\ X \end{pmatrix} &\le (1+q)^d \,\text{Vol}\begin{pmatrix} I \\ X \end{pmatrix} \\ \text{Vol}\begin{pmatrix} I+Y & X^T \\ X & -I \end{pmatrix} &\ge (1-q)^d \,\text{Vol}\begin{pmatrix} I & X^T \\ X & -I \end{pmatrix}. \end{aligned}$$

The proof is given in Appendix 6.2. Finally, let us recall several important geometric consequences involving the reach:

Proposition 9 *The following holds:*

i. For all $x, y \in \mathcal{M}$ such that $\|x-y\| \le \tau/2$, we have

$$d_{\mathcal{M}}(x,y) \le \tau - \tau\sqrt{1 - 2\frac{\|x-y\|}{\tau}} \le 2\|x-y\|.$$

ii. Let $\gamma(t) : [0,1] \mapsto \mathcal{M}$ be the arclength-parameterized geodesic. Then $\|\gamma''(t)\| \le \frac{1}{\tau}$ for all t.

iii. *Let ϕ be the angle between $T_x\mathcal{M}$ and $T_y\mathcal{M}$, in other words,*

$$\cos(\phi) := \min_{u\in T_x\mathcal{M}, \|u\|=1} \max_{v\in T_y\mathcal{M}, \|v\|=1} |\langle u, v\rangle| .$$

If $\|x-y\| \le \frac{\tau}{2}$, then $\cos(\phi) \ge \sqrt{1-2\frac{\|x-y\|}{\tau}}$.

iv. *If x is such that $\|x-y\| < \tau/2$, then x is a regular point of $\mathrm{Proj}_{y+T_y\mathcal{M}} : \mathcal{B}(y,\tau/2)\cap \mathcal{M} \to y+T_y\mathcal{M}$ (in other words, the Jacobian of $\mathrm{Proj}_{y+T_y\mathcal{M}}$ at x is nonsingular).*

v. *Let $y\in\mathcal{M}$, $r<\tau$ and $A = \mathcal{M}\cap B(y,r)$. Then*

$$B_d(y, r\cos(\theta)) \subseteq \mathrm{Proj}_{y+T_y\mathcal{M}}(A),$$

where $\theta = \arcsin\left(\frac{r}{2\tau}\right)$.

Proof Part *i.* is the statement of Proposition 6.3 and part *ii.* - of Proposition 6.1 in (Niyogi et al., 2008). Part *iii.* is demonstrated in Lemma 5.4 of the same paper, and this lemma coincides with *iv.* Part *v.* is proven in Lemma 5.3 of (Niyogi et al., 2008). ■

5. Proofs of the main results

The rest of the paper is devoted to the proofs of our main results.

5.1 Proof of Theorem 2

Assumption **(A3)** above controls the $L_2(\Pi)$ approximation error of $x\in M$ by $P_j(x)$ (see Remark 1, part *iii.*), hence we will concentrate on the stochastic error $\|\widehat{P}_j(x) - P_j(x)\|$. To this end, we will need to estimate $\|c_{j,k} - \widehat{c}_{j,k}\|$ and $\|\mathrm{Proj}_{V_{j,k}} - \mathrm{Proj}_{\widehat{V}_{j,k}}\|$, $k = 1\dots N(j)$.

One of the main tools required to obtain this bound is the noncommutative Bernstein's inequality.

Theorem 10 *(Minsker, 2013, Theorem 2.1) Let $Z_1,\dots,Z_n \in \mathbb{R}^{D\times D}$ be a sequence of independent symmetric random matrices such that $\mathbb{E}Z_i = 0$ and $\|Z_i\| \le U$ a.s., $1\le i\le n$. Let*

$$\sigma^2 := \left\| \sum_{i=1}^n \mathbb{E}Z_i^2 \right\|.$$

Then for any $t\ge 1$

$$\left\| \sum_{i=1}^n Z_i \right\| \le 2\max\left(\sigma\sqrt{t+\log(\bar{D})}, U(t+\log(\bar{D})) \right) \tag{14}$$

with probability $\ge 1-e^{-t}$, where $\bar{D} := 4\frac{\mathrm{tr}\left(\sum_{i=1}^n \mathbb{E}Z_i^2\right)}{\sigma^2}$.

Note that we always have $\bar{D} \le 4D$. We use this inequality to estimate $\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\|$: let $\Pi(dx|A)$ be the conditional distribution of X given that $X \in A$, and set $\Pi_{j,k}(dx) := \Pi(dx|C_{j,k})$. Let $m_{j,k} := \sum_{i=1}^{n} I\{X_i \in C_{j,k}\}$ to be the number of samples in $C_{j,k}$, $k = 1 \dots N(j)$. Let $I \subset \{1, \dots, n\}$ be such that $|I| = m$. Conditionally on the event $A_I := \{X_i \in C_{j,k} \text{ for } i \in I, \text{ and } X_i \notin C_{j,k} \text{ for } i \notin I\}$, the random variables $\{X_i,\ i \in I\}$ are independent with distribution $\Pi_{j,k}$. Then

$$\begin{aligned} \Pr\left(\left\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\right\| \ge s \,|\, m_{j,k} = m\right) &= \sum_{I \subset \{1,\dots,n\}, |I|=m} \Pr\left(\left\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\right\| \ge s \,|\, A_I\right) \frac{1}{\binom{n}{m}} \quad (15) \\ &= \Pr\left(\left\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\right\| \ge s \,|\, A_{\{1,\dots,m\}}\right). \end{aligned}$$

To estimate $\Pr\left(\left\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\right\| \ge s \,|\, A_{\{1,\dots,m\}}\right)$, we use the following inequality. Recall that

$$\bar{d} = 4d^2 \frac{\theta_2^4}{\theta_3^2},$$

where θ_2, θ_3 are the constants in Assumptions **(A2)** and **(A3)**.

Lemma 11 *Let $X, X_1, \dots, X_m$ be an i.i.d. sample from $\Pi_{j,k}$. Set*

$$\widehat{c}_{j,k} = \frac{1}{m} \sum_{i=1}^{m} X_i \quad \text{and} \quad \widehat{\Sigma}_{j,k} := \frac{1}{m} \sum_{i=1}^{m} (X_i - \widehat{c}_{j,k})(X_i - \widehat{c}_{j,k})^T.$$

Assume that $m \ge t + \log(\bar{d} \vee 8)$. Then with probability $\ge 1 - 2e^{-t}$,

$$\left\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\right\| \le 6r^2 \sqrt{\frac{t + \log(\bar{d} \vee 8)}{m}}.$$

Proof We want to estimate

$$\begin{aligned} \left\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\right\| &= \left\| \frac{1}{m} \sum_{i=1}^{m} (X_i - c_{j,k})(X_i - c_{j,k})^T - \Sigma_{j,k} + (c_{j,k} - \widehat{c}_{j,k})(c_{j,k} - \widehat{c}_{j,k})^T \right\| \\ &\le \left\| \frac{1}{m} \sum_{i=1}^{m} (X_i - c_{j,k})(X_i - c_{j,k})^T - \Sigma_{j,k} \right\| + \left\| (c_{j,k} - \widehat{c}_{j,k})(c_{j,k} - \widehat{c}_{j,k})^T \right\|. \quad (16) \end{aligned}$$

Set $r := \theta_2 \cdot 2^{-j}$. Recall that $\|x - c_{j,k}\| \le r$ for all $x, y \in C_{i,j}$ by assumption **(A2)**. It implies that

1. for all $1 \le i \le m$, $\|(X_i - c_{j,k})(X_i - c_{j,k})^T\| \le r^2$ almost surely,
2. $\left\|\mathbb{E}\left[(X_i - c_{j,k})(X_i - c_{j,k})^T\right]^2\right\| = \left\|\mathbb{E}\|X_i - c_{j,k}\|^2 (X_i - c_{j,k})(X_i - c_{j,k})^T\right\| \le r^2 \|\Sigma_{j,k}\|$.

Therefore, by Theorem 10 applied to $Z_i := \frac{1}{m}(X_i - c_{j,k})(X_i - c_{j,k})^T,\ i = 1 \dots m$,

$$\left\| \frac{1}{m} \sum_{i=1}^{m} (X_i - c_{j,k})(X_i - c_{j,k})^T - \Sigma_{j,k} \right\| \le 2 \left(r \sqrt{\frac{(t + \log(\bar{d}))\|\Sigma_{j,k}\|}{m}} \vee r^2 \frac{t + \log(\bar{d})}{m} \right)$$
$$= 2r^2 \sqrt{\frac{(t + \log(\bar{d}))}{m}} \left(\sqrt{\frac{t + \log(\bar{d})}{m}} \vee \sqrt{\left\| \frac{\Sigma_{j,k}}{r^2} \right\|} \right)$$

with probability $\ge 1 - e^{-t}$. Note that $\|\Sigma_{j,k}\| \le \operatorname{tr}(\Sigma_{j,k}) \le r^2$. Moreover,

$$\bar{D} = 4 \frac{\operatorname{tr}(\mathbb{E} Z_1^2)}{\|\mathbb{E} Z_1^2\|} \le 4 \frac{\mathbb{E}(\operatorname{tr} Z_1)^2}{\left(\lambda_d^{j,k}\right)^2} \le 4d^2 \frac{r^4}{\theta_3^2 2^{-4j}} = 4d^2 \frac{\theta_2^4}{\theta_3^2} = \bar{d}$$

by assumption **(A3)** and the definition of r. Since $\frac{t + \log(\bar{d})}{m} \le 1$ by assumption,

$$\left\| \frac{1}{m} \sum_{i=1}^{m} (X_i - c_{j,k})(X_i - c_{j,k}) - \Sigma_{j,k} \right\| \le 2r^2 \sqrt{\frac{t + \log(\bar{d})}{m}}.$$

For the second term in (16), note that $\|(c_{j,k} - \widehat{c}_{j,k})(c_{j,k} - \widehat{c}_{j,k})\| = \|c_{j,k} - \widehat{c}_{j,k}\|^2$. We apply Theorem 10 to the symmetric matrices

$$G_i := \begin{pmatrix} 0 & (X_i - c_{j,k})^T \\ X_i - c_{j,k} & 0 \end{pmatrix}.$$

Noting that $\|G_i\| = \|X_i - c_{j,k}\| \le r$ almost surely,

$$\|\mathbb{E} G_i^2\| = \mathbb{E}\|X_i - c_{j,k}\|^2 = \operatorname{tr}(\Sigma_{j,k}) \le r^2,$$

and $\frac{\operatorname{tr}(\mathbb{E} G_i^2)}{\|\mathbb{E} G_i^2\|} = 2$, we get that for all t such that $t + \log 8 \le m$, with probability $\ge 1 - e^{-t}$

$$\|\widehat{c}_{j,k} - c_{j,k}\| \le 2 \left[r \sqrt{\frac{(t + \log 8)}{m}} \vee r \frac{t + \log 8}{m} \right] \le 2r \sqrt{\frac{t + \log 8}{m}}, \tag{17}$$

hence with the same probability

$$\left\| (c_{j,k} - \widehat{c}_{j,k})(c_{j,k} - \widehat{c}_{j,k})^T \right\| \le 4r^2 \frac{t + \log 8}{m},$$

and the claim follows. ∎

Given the previous result, we can estimate the angle between the eigenspaces of $\widehat{\Sigma}_{j,k}$ and $\Sigma_{j,k}$:

Theorem 12 *(Davis and Kahan, 1970), or (Zwald and Blanchard, 2006, Theorem 3). Let $\delta_d = \delta_d(\Sigma_{j,k}) := \frac{1}{2}(\lambda_d^{j,k} - \lambda_{d+1}^{j,k})$. If $\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\| < \delta_d/2$, then*

$$\left\| \operatorname{Proj}_{V_{j,k}} - \operatorname{Proj}_{\widehat{V}_{j,k}} \right\| \le \frac{\|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\|}{\delta_d},$$

Since $\delta_d \geq \frac{\theta_3}{2\theta_2^2}\frac{r^2}{d}$ by assumption **(A3)**, the previous result implies that, conditionally on the event $\{m_{j,k} = m\}$, with probability $\geq 1 - 2e^{-t}$,

$$\left\| \operatorname{Proj}_{V_{j,k}} - \operatorname{Proj}_{\widehat{V}_{j,k}} \right\| \leq 12d\frac{\theta_2^2}{\theta_3}\sqrt{\frac{t + \log(\bar{d} \vee 8)}{m}}.$$

It remains to obtain the unconditional bound. Set $n_{j,k} := n\Pi(C_{j,k})$ and note that $n_{j,k} \geq \theta_1 n 2^{-jd}$ by assumption **(A1)**. To this end, we have

$$\begin{aligned}
&\Pr\left(\max_{k=1\ldots N(j)} \left\| \operatorname{Proj}_{V_{j,k}} - \operatorname{Proj}_{\widehat{V}_{j,k}} \right\| \geq 12\frac{\theta_2^2}{\theta_3}\sqrt{\frac{(t + \log(\bar{d} \vee 8))d^2}{n_{j,k}/2}} \right) \\
&\leq \Pr\left(\max_{k=1\ldots N(j)} \left\| \operatorname{Proj}_{V_{j,k}} - \operatorname{Proj}_{\widehat{V}_{j,k}} \right\| \geq 12\frac{\theta_2^2}{\theta_3}\sqrt{\frac{(t + \log(\bar{d} \vee 8))d^2}{n_{j,k}/2}} \middle| m_{j,k} \geq n_{j,k}/2,\ k = 1\ldots N(j) \right) \\
&+ \Pr\left(\bigcup_{k=1}^{N(j)} \{m_{j,k} < n_{j,k}/2\} \right) \leq N(j)e^{-t} + \sum_{k=1}^{N(j)} \Pr\left(m_{j,k} < n_{j,k}/2\right).
\end{aligned}$$

Recall that $m_{j,k} = \sum_{i=1}^{n} I\{X_i \in C_{j,k}\}$, hence $\mathbb{E}m_{j,k} = n_{j,k}$ and $\operatorname{Var}(m_{j,k}) \leq n_{j,k}$. Bernstein's inequality (see Lemma 2.2.9 in van der Vaart and Wellner, 1996) implies that

$$|m_{j,k} - n_{j,k}| \leq \left(2\sqrt{sn_{j,k}} \vee \frac{4}{3}s \right)$$

with probability $\geq 1 - e^{-s}$. Choosing $s = \frac{n_{j,k}}{16}$, we deduce that $\Pr\left(m_{j,k} < n_{j,k}/2\right) \leq e^{-\frac{\theta_1}{16}n2^{-jd}}$, and, since $N(j) \leq \frac{1}{\theta_1}2^{jd}$ by assumption **(A1)**,

$$\sum_{k=1}^{N(j)} \Pr\left(m_{j,k} < n_{j,k}/2\right) \leq \frac{1}{\theta_1}2^{jd}e^{-\frac{\theta_1}{16}n2^{-jd}}$$

and

$$\Pr\left(\max_{k=1\ldots N(j)} \left\| \operatorname{Proj}_{V_{j,k}} - \operatorname{Proj}_{\widehat{V}_{j,k}} \right\| \geq 12\frac{\theta_2^2}{\theta_3}\sqrt{\frac{(t + \log(\bar{d} \vee 8))d^2}{n_{j,k}/2}} \right) \leq \frac{2^{jd}}{\theta_1}\left(e^{-t} + e^{-\frac{\theta_1}{16}n2^{-jd}} \right) \tag{18}$$

A similar argument implies that

$$\Pr\left(\max_{k=1\ldots N(j)} \|c_{j,k} - \widehat{c}_{j,k}\| \geq 2r\sqrt{\frac{t + \log(\bar{d} \vee 8)}{n_{j,k}/2}} \right) \leq \frac{2^{jd}}{\theta_1}\left(e^{-t} + e^{-\frac{\theta_1}{16}n2^{-jd}} \right). \tag{19}$$

We are in position to conclude the proof of Theorem 2: given $x \in C_{j,k}$, note that

$$\begin{aligned}
\|P_j(x) - \widehat{P}_j(x)\| &= \|c_{j,k} - \widehat{c}_{j,k} + \operatorname{Proj}_{V_{j,k}}(x - c_{j,k}) - \operatorname{Proj}_{\widehat{V}_{j,k}}(x - c_{j,k} + c_{j,k} - \widehat{c}_{j,k})\| \\
&\leq 2\|c_{j,k} - \widehat{c}_{j,k}\| + \|\operatorname{Proj}_{V_{j,k}} - \operatorname{Proj}_{\widehat{V}_{j,k}}\| \cdot \|x - c_{j,k}\|.
\end{aligned}$$

Together with assumption **(A2)**, (18) and (19), it implies that with high probability

$$\|P_j(x) - \widehat{P}_j(x)\| \le 4\sqrt{2}\frac{\theta_2}{\sqrt{\theta_1}}2^{-j}\sqrt{\frac{t+\log(\bar{d}\vee 8)}{n2^{-jd}}} + 12\sqrt{2}\frac{\theta_2^3}{\theta_3\sqrt{\theta_1}}2^{-j}\sqrt{\frac{(t+\log(\bar{d}\vee 8))d^2}{n2^{-jd}}}.$$

Combined with assumption **(A3)** (see Remark 1, part *iii.*), this yields the result.

5.2 Proof of Theorem 6

Recall that $\mathcal{M} \hookrightarrow \mathbb{R}^D$ is a smooth (or at least C^2) compact manifold without boundary, with reach τ, and equipped with the volume measure $d\mathrm{Vol}_{\mathcal{M}}$. Our proof is divided into several steps, and each of them is presented in a separate subsection to improve readability.

5.2.1 Local inversions of the projection

In this section, we show that, for $r < \tau/8$, the projection map $\mathrm{Proj}_{y+T_y\mathcal{M}}$ is injective on $B(y,r)\cap\mathcal{M}$, and hence invertible by part *iv.* of Proposition 9. We also demonstrate that the derivatives of this inverse are bounded in a suitable sense. These estimates shall allow us to develop bounds on volumes in $\mathcal{M}_\sigma$.

We begin by proving a bound on the local deviation of the manifold from a tangent plane.

Lemma 13 *Suppose $\eta \in T_y^{\perp}\mathcal{M}$ with $\|\eta\| = 1$ and $z \in B(y,r)\cap\mathcal{M}$, where $r \le \tau/2$. Then*

$$|\langle \eta, z-y\rangle| \le \frac{2r^2}{\tau}$$

Proof Let $\gamma : [0, d_{\mathcal{M}}(z,y)] \to \mathcal{M}$ denote the arclength-parameterized geodesic connecting y to z in $\mathcal{M}$. Since γ is a geodesic, there is a $v \in T_y\mathcal{M}$ with $\|v\| = 1$ such that the Taylor expansion

$$z = y + d_{\mathcal{M}}(z,y)v + \int_0^{d_{\mathcal{M}}(z,y)} \gamma''(t)\,(d_{\mathcal{M}}(z,y) - t)\,dt.$$

By Proposition 9, $\|\gamma''(t)\|_2 \le 1/\tau$ for all t and $d_{\mathcal{M}}(z,y) \le 2r$, so we have that

$$\begin{aligned}
|\langle \eta, z-y\rangle| &= \left|\left\langle \eta, \int_0^{d_{\mathcal{M}}(z,y)} \gamma''(t)\,(d_{\mathcal{M}}(z,y)-t)\,dt\right\rangle\right| \\
&\le \int_0^{d_{\mathcal{M}}(z,y)} |\langle \eta, \gamma''(t)\rangle|\,(d_{\mathcal{M}}(z,y)-t)\,dt \\
&\le \frac{1}{\tau}\int_0^{d_{\mathcal{M}}(z,y)} (d_{\mathcal{M}}(z,y)-t)\,dt \\
&\le \frac{d_{\mathcal{M}}(z,y)^2}{2\tau} \\
&\le \frac{2r^2}{\tau}.
\end{aligned}$$

■

Our next lemma quantitatively establishes the local injectivity of the affine projections onto tangent spaces. [1]

Lemma 14 *Suppose $y \in \mathcal{M}$ and $r < \tau/8$. Then $\text{Proj}_{y+T_y\mathcal{M}} : B(y,r) \cap \mathcal{M} \to y + T_y\mathcal{M}$ is injective.*

Proof Suppose a and b are distinct in $B(y,r) \cap \mathcal{M}$. Now, $b - a = v + w$ where $v \in T_a\mathcal{M}$ and $w \in T_a^{\perp}\mathcal{M}$, and note that $\|w\| \leq \frac{2\|b-a\|^2}{\tau} \leq 4\frac{r}{\tau}$ by Lemma 13. This also implies that

$$\|v\| = \sqrt{\|a-b\|^2 - \|w\|^2} \geq \sqrt{\|a-b\|^2 - 4\frac{\|a-b\|^4}{\tau^2}} \geq \|a-b\|\sqrt{1 - 16\frac{r^2}{\tau^2}} \geq \|a-b\|\sqrt{1-4\frac{r}{\tau}}\,.$$

By part *iii.* of Proposition 9, there is a $u \in T_y\mathcal{M}$ such that $\langle u, v\rangle \geq \|v\|\cos(\phi)$ where ϕ is the angle between $T_y\mathcal{M}$ and $T_a\mathcal{M}$. Then

$$\begin{aligned}
|\langle u, b-a\rangle| &\geq |\langle u, v\rangle| - |\langle u, w\rangle| \\
&\geq \|v\|\cos(\phi) - \|w\| \\
&\geq \|a-b\|\sqrt{1-4\frac{r}{\tau}}\sqrt{1-2\frac{r}{\tau}} - 2\frac{\|a-b\|^2}{\tau} \\
&\geq \|b-a\|\left(\sqrt{1-4\frac{r}{\tau}}\sqrt{1-4\frac{r}{\tau}} - 4\frac{r}{\tau}\right) \\
&\geq \|b-a\|\left(1 - 8\frac{r}{\tau}\right).
\end{aligned}$$

It then follows from $r < \tau/8$ that $\text{Proj}_{T_y\mathcal{M}}(b-a) \neq 0$, and hence $\text{Proj}_{y+T_y\mathcal{M}}(a) \neq \text{Proj}_{y+T_y\mathcal{M}}(b)$ and injectivity holds. ■

There are two important conclusions that Lemma 14 provides. First of all, it indicates that, under a certain radius bound, the manifold does not "curve back" into particular regions. This is helpful when we begin to examine upper bounds on local volumes. More importantly, if we let $J_{y,r} = \text{Proj}_{y+T_y\mathcal{M}}(B(y,r) \cap \mathcal{M})$, then there is a well-defined inverse map f of $\text{Proj}_{y+T_y\mathcal{M}}$, $f : J_{y,r} \to B(y,r) \cap \mathcal{M}$, when $r < \tau/8$. Part *iv* of Proposition 9 implies that f is at least a C^2 function, and part *v* of Proposition 9 implies that there is a d-dimensional ball inside of $J_{y,r}$ of radius $\cos(\theta)r$, where $\theta = \arcsin(r/2\tau)$.

Whenever we refer to such an f, we think of $J_{y,r}$ as a subset in the span of the first d canonical directions, and we identify f with the value f takes in the span of the remaining $D-d$ directions. Thus, we identify f with the function whose graph is a small part of the manifold. Such an identification is obtained via an affine transformation, so we may do this without any loss of generality. Using these assumptions, we may prove the following bounds.

1. In an independent work, Eftekhari and Wakin (2013) prove a slightly stronger result that holds for $r < \tau/4$.

Proposition 15 *Let $\varepsilon < \tau/8$, and assume f is defined above so that $v \longmapsto \begin{pmatrix} v \\ f(v) \end{pmatrix}$ is the inverse of $\mathrm{Proj}_{y+T_y\mathcal{M}}$ in $B(y,\varepsilon)$ for some $y \in \mathcal{M}$. Then*

$$\sup_{v \in B_d(0,\varepsilon)} \|Df(v)\| \leq \frac{2\varepsilon}{\tau - 2\varepsilon} \tag{20}$$

and

$$\sup_{v \in B_d(0,\varepsilon)} \sup_{u \in \mathcal{S}^{D-d-1}} \left\| \sum_{i=1}^{D-d-1} u_i D^2 f_i(v) \right\| \leq \frac{\tau^2}{(\tau - 2\varepsilon)^3}. \tag{21}$$

Proof For $\varepsilon < \tau/8$, we may define the embedding

$$\begin{pmatrix} v \\ \beta \end{pmatrix} \longmapsto \begin{pmatrix} v \\ f(v) \end{pmatrix} + \begin{pmatrix} Df(v)^T \\ -I \end{pmatrix} \beta$$

where we have assumed (without loss of generality) that $y = 0$ and $T_y\mathcal{M}$ coincides with the span of the first d canonical orthonormal basis members. The domain of this map is the set

$$\Omega = \{(v,\beta) \in \mathbb{R}^d \times \mathbb{R}^{D-d} : v \in T_y\mathcal{M} \cap B(0,\varepsilon), \|\beta\|^2 + \|Df(v)^T\beta\|^2 < \tau^2\}$$

and the Jacobian of this map is

$$\begin{pmatrix} I + \sum_{i=1}^{D-d} \beta_i D^2 f_i(v) & Df(v)^T \\ Df(v) & -I \end{pmatrix}.$$

It is clear that the inverse of the above map is given by

$$x \longmapsto (\mathrm{Proj}_{y+T_y\mathcal{M}}(\mathrm{Proj}_{\mathcal{M}}(x)), \mathrm{Proj}_{T_y^\perp\mathcal{M}}(x - \mathrm{Proj}_{\mathcal{M}}(x))),$$

which is at least a C^1 map. Thus, a necessary condition for the τ-radius normal bundle to embed is that the Jacobian exhibited above is invertible, which in turn implies that

$$\begin{pmatrix} I + \sum_{i=1}^{D-d} \beta_i D^2 f_i(v) & Df(v)^T \\ Df(v) & -I \end{pmatrix} \begin{pmatrix} \zeta \\ Df(v)\zeta \end{pmatrix} \neq 0$$

for all $\zeta \neq 0$ when $(v,\beta) \in \Omega$. This reduces to $(I + \sum \beta_i D^2 f_i(v) + Df(v)^T Df(v))\zeta \neq 0$, and so a necessary condition for embedding is then that the norm of $\sum_{i=1}^{D-d} \beta_i D^2 f_i(v)$ does not exceed $1 + \|Df(v)\|^2$ whenever

$$\left\| \begin{pmatrix} Df(v)^T \\ -I \end{pmatrix} \beta \right\|^2 = \|\beta\|^2 + \|Df(v)^T\beta\|^2 < \tau^2.$$

In particular, this must be true if $\|\beta\| < \tau/\sqrt{1 + \|Df(v)\|^2}$. This reduces to the condition that the operator norm

$$\sup_{u \in \mathcal{S}^{D-d-1}} \left\| \sum_{i=1}^{D-d} u_i D^2 f_i(v) \right\| < \frac{(1 + \|Df(v)\|^2)^{3/2}}{\tau} < \frac{1}{\tau}\left(1 + \|Df(v)\|\right)^3. \tag{22}$$

By the fundamental theorem of calculus, we have that

$$Df(v)x = Df(0)x + \int_0^{\|v\|} [u_v^T D^2 f_i(tu_v)x] dt = \int_0^{\|v\|} [u_v^T D^2 f_i(tu_v)x] dt,$$

where $u_v = v/\|v\|$ and $[u_v^T D^2 f_i(tu_v)x]$ indicates a vector with ith component $u_v^T D^2 f_i(tu_v)x$. Consequently, for any $x \in \mathbb{R}^d$, we have that

$$\begin{aligned} \|Df(v)x\| &\leq \int_0^{\|v\|} \left\| [u_v^T D^2 f_i(tu_v)x] \right\| dt \leq \|v\| \sup_{t\in[0,\|v\|]} \left\| [u_v^T D^2 f_i(tu_v)x] \right\| \\ &\leq \varepsilon \sup_{t\in[0,\varepsilon]} \left\| [u_v^T D^2 f_i(tu_v)x] \right\|. \end{aligned} \tag{23}$$

Now,

$$\begin{aligned} \left\| [u_v^T D^2 f_i(tu_v)x] \right\| &= \sup_{u\in\mathcal{S}^{D-d-1}} \langle u, [u_v^T D^2 f_i(tu_v)x] \rangle = \sup_{u\in\mathcal{S}^{D-d-1}} \sum_{i=1}^{D-d} u_i (u_v^T D^2 f_i(tu_v)x) \\ &= \sup_{u\in\mathcal{S}^{D-d-1}} u_v^T \left(\sum_{i=1}^{D-d} u_i D^2 f_i(tu_v) \right) x \\ &\leq \sup_{u\in\mathcal{S}^{D-d-1}} \|u_v\| \left\| \sum_{i=1}^{D-d} u_i D^2 f_i(tu_v) \right\| \|x\| \\ &= \|x\| \sup_{u\in\mathcal{S}^{D-d-1}} \left\| \sum_{i=1}^{D-d} u_i D^2 f_i(tu_v) \right\|, \end{aligned}$$

which together with (23) and (22) yields the bound

$$\|Df(v)\| < \frac{\varepsilon}{\tau} \left(1 + \sup_{t\in[0,\varepsilon]} \|Df(tu_v)\| \right)^3.$$

Since this inequality also holds for any v' with $\|v\| \leq \varepsilon'$, taking a supremum yields

$$\sup_{\varepsilon'\in[0,\varepsilon]} \|Df(tu_v)\| \leq \sup_{\varepsilon'\in[0,\varepsilon]} \frac{\varepsilon'}{\tau} \left(1 + \sup_{t\in[0,\varepsilon']} \|Df(tu_v)\| \right)^3 \leq \frac{\varepsilon}{\tau} \left(1 + \sup_{\varepsilon'\in[0,\varepsilon]} \|Df(tu_v)\| \right)^3,$$

and hence

$$\sup_{v\in B_d(0,\varepsilon)} \|Df(v)\| \leq \frac{\varepsilon}{\tau} \left(1 + \sup_{v\in B_d(0,\varepsilon)} \|Df(v)\| \right)^3.$$

Setting $a(\varepsilon') = \sup_{v\in B_d(0,\varepsilon')} \|Df(v)\|$, we have that $a(0) = 0$,

$$a(\varepsilon') \leq \frac{\varepsilon'}{\tau} \left(1 + a(\varepsilon') \right)^3,$$

for all $\varepsilon' \geq 0$, and a is continuous by continuity of $\|Df(v)\|$. Setting $b(\varepsilon') = a(\varepsilon')/(1+a(\varepsilon'))$, we get

$$b(\varepsilon')(1 - b(\varepsilon'))^2 \leq \frac{\varepsilon'}{\tau}.$$

Examining the polynomial $x(1-x)^2$, we see that the sublevel set $x(1-x)^2 \leq \omega$ consists of two components when $\omega < 4/27$. Also note that if $\omega < 1/8$, then

$$2(1-2\omega)^2 = 2 - 8\omega + 8\omega^2 > 2 - 1 = 1,$$

and hence

$$2\omega(1-2\omega)^2 > \omega.$$

Consequently, if x is such that $x(1-x)^2 \leq \omega$ and is in the interval containing zero in the sublevel set $x(1-x)^2 \leq \omega < 1/8$, then $x \leq 2\omega$.

By these observations, continuity of $b(\varepsilon')$, and the fact that $b(0) = 0$, we have that $a(\varepsilon') \leq \frac{2\frac{\varepsilon'}{\tau}}{1-2\frac{\varepsilon'}{\tau}}$, and thus

$$\sup_{v \in B_d(0,\varepsilon)} \|Df(v)\| \leq \frac{2\varepsilon}{\tau - 2\varepsilon}.$$

From the bound in (22) we now acquire the bound

$$\sup_{v \in B_d(0,\varepsilon)} \sup_{u \in \mathcal{S}^{D-d-1}} \left\| \sum_{i=1}^{D-d-1} u_i D^2 f_i(v) \right\| \leq \frac{\tau^2}{(\tau - 2\varepsilon)^3}.$$

■

5.2.2 Volume bounds

The main result of this section is Lemma 16, which allows us to compare volumes in $\mathcal{M}_\sigma$ with volumes in $\mathcal{M}$. It also establishes an upper bound on volumes, which is an essential ingredient when we control the conditional distribution of Π subject to being in a particular $C_{j,k}$.

Lemma 16 *Suppose $\sigma < \tau$, suppose $U \subseteq \mathcal{M}$ is measurable, and define $P : \mathcal{M}_\sigma \to \mathcal{M}$ so that $x \mapsto \text{Proj}_{\mathcal{M}}(x)$ under P. Then*

i.

$$\left(1 - \frac{\sigma}{\tau}\right)^d \text{Vol}_{\mathcal{M}}(U)\text{Vol}(B_{D-d}(0,\sigma)) \leq \text{Vol}(P^{-1}(U)) \leq \left(1 + \frac{\sigma}{\tau}\right)^d \text{Vol}_{\mathcal{M}}(U)\text{Vol}(B_{D-d}(0,\sigma))$$

ii. If $r + \sigma \leq \tau/8$, then

$$\text{Vol}(\mathcal{M}_\sigma \cap B(y,r)) \leq \left(1 + \frac{\sigma}{\tau}\right)^d \left(1 + \left(\frac{2(r+\sigma)}{\tau - 2(r+\sigma)}\right)^2\right)^{d/2} \text{Vol}(B_d(0, r+\sigma))\text{Vol}(B_{D-d}(0,\sigma)).$$

Proof We first prove part *i*. Let $\varepsilon > 0$ satisfy $\varepsilon < \tau/8$. Because of (21) and the fact that $\|\beta\| \leq \sigma$, we have that

$$\left\| \sum_{i=1}^{D-d} \beta_i D^2 f_i(v) \right\| \leq \frac{\sigma\tau^2}{(\tau - 2\varepsilon)^3}.$$

Since this is also a bound for the columns of $\sum \beta_i D^2 f_i(v)$, Proposition 8 implies that

$$\mathrm{Vol}\begin{pmatrix} I + \sum \beta_i D^2 f_i(v) & Df(v)^T \\ Df(v) & -I \end{pmatrix} \leq \left(1 + \frac{\sigma\tau^2}{(\tau - 2\varepsilon)^3}\right)^d \mathrm{Vol}\begin{pmatrix} I & Df(v)^T \\ Df(v) & -I \end{pmatrix}$$

in $T^{\perp}(\mathcal{M} \cap B(y, \varepsilon)) \cap \mathcal{M}_{\sigma}$.

On the other hand, we have that

$$\mathrm{Vol}\begin{pmatrix} Df^T(v) \\ -I \end{pmatrix} \leq \prod_{i=1}^{D-d} \sqrt{1 + \|\nabla f_i(v)\|^2} \leq \left(1 + \frac{4\varepsilon^2}{(\tau - 2\varepsilon)^2}\right)^{(D-d)/2}$$

since (20) implies the bounds $\|\frac{\partial f(v)}{\partial v_i}\| \leq \frac{2\varepsilon}{\tau - 2\varepsilon}$ for each $i = 1, \ldots, d$, and the above is the largest this quantity may be subject to these bounds.

When these estimates are joined together, we have an inequality

$$\begin{aligned}\mathrm{Vol}\begin{pmatrix} I + \sum_{i=1}^{D-d} \beta_i D^2 f_i(v) & Df(v)^T \\ Df(v) & -I \end{pmatrix} &\leq \left(1 + \frac{\sigma\tau^2}{(\tau - 2\varepsilon)^3}\right)^d \mathrm{Vol}\begin{pmatrix} I & Df(v)^T \\ Df(v) & -I \end{pmatrix} \\ &\leq \left(1 + \frac{\sigma\tau^2}{(\tau - 2\varepsilon)^3}\right)^d \left(1 + \frac{4\varepsilon^2}{(\tau - 2\varepsilon)^2}\right)^{(D-d)/2} \mathrm{Vol}\begin{pmatrix} I \\ Df(v) \end{pmatrix}.\end{aligned}$$

For an arbitrarily small $\varepsilon > 0$, let $\{U_\gamma\}_{\gamma \in \Gamma}$ denote a finite partition of U into measurable sets such that there for each $\gamma \in \Gamma$, there is a y_γ satisfying $U_\gamma \subset \mathcal{M} \cap B(y_\gamma, \varepsilon)$. Let f_γ denote the inverse of $P_\gamma = \mathrm{Proj}_{\, y_\gamma + T_{y_\gamma}\mathcal{M}}$ in U_γ, and set

$$E_{\gamma,v} = \{\beta \in \mathbb{R}^{D-d} : \|\beta\|^2 + \|Df_\gamma(v)\beta\|^2 \leq \sigma^2\}$$

for all $v \in P_\gamma(U_\gamma)$. Thus,

$$\begin{aligned}\int_{P_\gamma^{-1}(U_\gamma)} d\mathrm{Vol}(x) &= \int_{P_\gamma(U_\gamma)} \int_{E_{\gamma,v}} \mathrm{Vol}\begin{pmatrix} I + \sum_{i=1}^{D-d} \beta_i D^2 f_i(v) & Df(x)^T \\ Df(v) & -I \end{pmatrix} d\beta dv \\ &\leq \int_{P_\gamma(U_\gamma)} \int_{E_{\gamma,v}} \left(1 + \frac{\sigma\tau^2}{(\tau - 2\varepsilon)^3}\right)^d \left(1 + \frac{4\varepsilon^2}{(\tau - 2\varepsilon)^2}\right)^{d/2} \mathrm{Vol}\begin{pmatrix} I \\ Df(v) \end{pmatrix} d\beta dv \\ &\leq \left(1 + \frac{\sigma\tau^2}{(\tau - 2\varepsilon)^2}\right)^d \left(1 + \frac{4\varepsilon^2}{(\tau - 2\varepsilon)^2}\right)^{d/2} \mathrm{Vol}_{\,\mathcal{M}}(U_\gamma)\mathrm{Vol}(B_{D-d}(0, \sigma))\end{aligned}$$

since $E_{\gamma,v} \subset B_{D-d}(0, \sigma)$. Consequently, we have that

$$\begin{aligned}\mathrm{Vol}(P^{-1}(U)) &= \sum_{\gamma \in \Gamma} \mathrm{Vol}(P_\gamma^{-1}(U_\gamma)) \\ &\leq \sum_{\gamma \in \Gamma} \left(1 + \frac{\sigma\tau^2}{(\tau - 2\varepsilon)^3}\right)^d \left(1 + \frac{4\varepsilon^2}{(\tau - 2\varepsilon)^2}\right)^{d/2} \mathrm{Vol}_{\,\mathcal{M}}(U_\gamma)\mathrm{Vol}(B_{D-d}(0, \sigma)) \\ &= \left(1 + \frac{\sigma\tau^2}{(\tau - 2\varepsilon)^3}\right)^d \left(1 + \frac{4\varepsilon^2}{(\tau - 2\varepsilon)^2}\right)^{d/2} \mathrm{Vol}_{\,\mathcal{M}}(U)\mathrm{Vol}(B_{D-d}(0, \sigma)).\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, we obtain

$$\mathrm{Vol}(P^{-1}(U)) \cap \mathcal{M}_\sigma) \leq \left(1 + \frac{\sigma}{\tau}\right)^d \mathrm{Vol}_{\mathcal{M}}(U)\mathrm{Vol}(B_{D-d}(0,\sigma)).$$

This completes the proof of upper bound in part *i*. Using a similar partition strategy, we have that

$$\begin{aligned}
\int_{P_\gamma^{-1}(U_\gamma)} d\mathrm{Vol}(x) &= \int_{P_\gamma(U_\gamma)} \int_{E_{\gamma,v}} \mathrm{Vol}\begin{pmatrix} I + \sum_{i=1}^{D-d} \beta_i D^2 f_i(v) & Df(x)^T \\ Df(v) & -I \end{pmatrix} d\beta dv \\
&\geq \int_{P_\gamma(U_\gamma)} \int_{E_{\gamma,v}} \left(1 - \frac{\sigma\tau^2}{(\tau - 2\varepsilon)^3}\right)^d \mathrm{Vol}\begin{pmatrix} I & Df(v)^T \\ Df(v) & -I \end{pmatrix} d\beta dv \\
&= \int_{P_\gamma(U_\gamma)} \int_{E_{\gamma,v}} \left(1 - \frac{\sigma\tau^2}{(\tau - 2\varepsilon)^3}\right)^d \mathrm{Vol}\begin{pmatrix} I \\ Df(v) \end{pmatrix} \mathrm{Vol}\begin{pmatrix} Df(v)^T \\ -I \end{pmatrix} d\beta dv \\
&\geq \int_{P_\gamma(U_\gamma)} \int_{E_{\gamma,v}} \left(1 - \frac{\sigma\tau^2}{(\tau - 2\varepsilon)^3}\right)^d \mathrm{Vol}\begin{pmatrix} I \\ Df(v) \end{pmatrix} d\beta dv \\
&\geq \int_{P_\gamma(U_\gamma)} \int_{B_{D-d}\left(0, \frac{\sigma}{1 + \frac{\varepsilon}{\tau - \varepsilon}}\right)} \left(1 - \frac{\sigma\tau^2}{(\tau - 2\varepsilon)^3}\right)^d \mathrm{Vol}\begin{pmatrix} I \\ Df(v) \end{pmatrix} d\beta dv \\
&= \left(1 - \frac{\sigma\tau^2}{(\tau - 2\varepsilon)^3}\right)^d \mathrm{Vol}_{\mathcal{M}}(U_\gamma)\mathrm{Vol}\left(B_{D-d}\left(0, \left(1 - \frac{\varepsilon}{\tau}\right)\sigma\right)\right)
\end{aligned}$$

In the inequalities above, we have used the fact that there is a ball of radius $\left(1 - \frac{\varepsilon}{\tau}\right)\sigma$ inside of $E_{\gamma,v}$ for each γ and each v. Aggregating all of the sums and letting $\varepsilon \to 0$ yields the lower bound in part *i*.

We now prove part *ii*. Note that

$$\mathrm{Vol}(\mathcal{M}_\sigma \cap B(y,r)) \leq \mathrm{Vol}(P^{-1}(\mathcal{M} \cap B(y, r+\sigma)))$$

since $\|\mathrm{Proj}_{\mathcal{M}}(x) - y\| \leq \|x - y\| + \|\mathrm{Proj}_{\mathcal{M}}(x) - x\| \leq r + \sigma$. Part *ii*. now follows from part *i*. and the fact that

$$\begin{aligned}
\mathrm{Vol}_{\mathcal{M}}(\mathcal{M} \cap B(y, r+\sigma)) &\leq \int\limits_{P(\mathcal{M} \cap B(y, r+\sigma))} \mathrm{Vol}\begin{pmatrix} I \\ Df(v) \end{pmatrix} dv \\
&\leq \left(1 + \left(\frac{2(r+\sigma)}{\tau - 2(r+\sigma)}\right)^2\right)^{d/2} \mathrm{Vol}(B_d(0, r+\sigma)).
\end{aligned}$$

■

5.2.3 Absolute continuity of the pushforward of $U_{\mathcal{M}_\sigma}$ and local moments

Recall that $U_{\mathcal{M}_\sigma}$ is the uniform distribution over $\mathcal{M}_\sigma$, and let $U_{\mathcal{M}} := \frac{d\,\mathrm{Vol}_{\mathcal{M}}}{\mathrm{Vol}_{\mathcal{M}}(\mathcal{M})}$ be the uniform distribution over $\mathcal{M}$. In this section, we exploit the volume bounds of the previous

subsection to obtain control over probabilities and local moments of $U_{\mathcal{M}_\sigma}$. Our first result allows us to get the lower bounds for the "small ball" probabilities associated to $U_{\mathcal{M}_\sigma}$ that are independent of the ambient dimension D.

Lemma 17 *Suppose $\sigma < \tau$, and let $\widetilde{U}_{\mathcal{M}_\sigma}$ denote the pushforward of $U_{\mathcal{M}_\sigma}$ under $\operatorname{Proj}_{\mathcal{M}}$. Then $\widetilde{U}_{\mathcal{M}_\sigma}$ and $U_{\mathcal{M}}$ are mutually absolutely continuous with respect to each other, and*

$$\left(\frac{\tau-\sigma}{\tau+\sigma}\right)^d \leq \frac{d\widetilde{U}_{\mathcal{M}_\sigma}}{dU_{\mathcal{M}}} \leq \left(\frac{\tau+\sigma}{\tau-\sigma}\right)^d.$$

Proof This is a straightforward consequence of part *i.* of Lemma 16. ■

Lemma 18 *Suppose Π is a distribution supported on $\mathcal{M}_\sigma$, and let $r < \tau/2$. Further assume that Z is the random variable drawn from Π conditioned on the event $Z \in Q$ where $\mathcal{M}_\sigma \cap Q \subset B(y,r)$ for some $y \in \mathcal{M}$. If Σ is the covariance matrix of Z, then*

$$\sum_{i=d+1}^{D} \lambda_i(\Sigma) \leq 2\sigma^2 + \frac{8r^4}{\tau^2},$$

where $\lambda_i(\Sigma)$ are the eigenvalues of Σ arranged in the decreasing order.

Proof By the variational characterization of eigenvalues, we have that

$$\begin{aligned}\sum_{i=d+1}^{D} \lambda_i(\Sigma) &= \operatorname*{argmin}_{\dim(V)=D-d} \operatorname{tr}\left(\operatorname{Proj}_V^T \Sigma \operatorname{Proj}_V\right) \\ &= \operatorname*{argmin}_{\dim(V)=D-d} \mathbb{E}\|\operatorname{Proj}_V(Z-\mathbb{E}Z)\|^2 \\ &= \operatorname*{argmin}_{\dim(V)=d} \mathbb{E}\|Z-\mathbb{E}Z-\operatorname{Proj}_V(Z-\mathbb{E}Z)\|^2.\end{aligned}$$

Thus, we have that $\sum_{i=d+1}^{D} \lambda_i(\Sigma) \leq \mathbb{E}\|Z-\mathbb{E}Z-\operatorname{Proj}_{T_y\mathcal{M}}(Z-\mathbb{E}Z)\|^2$. Observe that

$$\begin{aligned}\mathbb{E}\|Z-\mathbb{E}Z-\operatorname{Proj}_{T_y\mathcal{M}}(Z-\mathbb{E}Z)\|^2 =& \mathbb{E}\|Z-y+(y-\mathbb{E}Z)-\operatorname{Proj}_{T_y\mathcal{M}}((Z-y)+(y-\mathbb{E}Z))\|^2 \\ =& \mathbb{E}\|Z-y-\operatorname{Proj}_{T_y\mathcal{M}}(Z-y)\|^2 \\ &- \|(y-\mathbb{E}Z)-\operatorname{Proj}_{T_y\mathcal{M}}(y-\mathbb{E}Z)\|^2 \\ &\leq \mathbb{E}\|Z-y-\operatorname{Proj}_{T_y\mathcal{M}}(Z-y)\|^2.\end{aligned}$$

Now for any $z \in \mathcal{M}_\sigma \cap B(y,r)$, we have that $z = \beta + x$ where $x \in \mathcal{M}$, and $\beta \in T_x^\perp \mathcal{M}$ satisfies $\|\beta\| \leq \sigma$. Moreover, there is a unique decomposition $x = \eta + v + y$ where $\eta \in T_y^\perp \mathcal{M}$ and $v \in T_y\mathcal{M}$. Thus,

$$\|z-y-\operatorname{Proj}_{T_y\mathcal{M}}(z-y)\| = \|\beta+\eta-\operatorname{Proj}_{T_y\mathcal{M}}\beta\| \leq \|\beta-\operatorname{Proj}_{T_y\mathcal{M}}(\beta)\| + \|\eta\| \leq \sigma + \frac{2r^2}{\tau}, \tag{24}$$

by Lemma 13, and we obtain the bound

$$\mathbb{E}\|Z - \mathbb{E}Z - \mathrm{Proj}_{T_y\mathcal{M}}(Z - \mathbb{E}Z)\|^2 \leq 2\sigma^2 + \frac{8r^4}{\tau^2}. \tag{25}$$

This establishes the required estimate. ■

Finally, we derive a lower bound on the upper eigenvalues of the local covariance for the uniform distribution (needed to satisfy assumption **(A3)**). This is done in the following lemma.

Lemma 19 *Suppose that $Q \subseteq \mathbb{R}^D$ is such that*

$$B(y, r_1) \subseteq Q \text{ and } \mathcal{M}_\sigma \cap Q \subset B(y, r_2)$$

for some $y \in \mathcal{M}$ and $\sigma < r_1 < r_2 < \tau/8 - \sigma$. Let Z be drawn from $U_{\mathcal{M}_\sigma}$ conditioned on the event $Z \in Q$, and suppose Σ is the covariance matrix of Z. Then

$$\lambda_d(\Sigma) \geq \frac{1}{4\left(1 + \frac{\sigma}{\tau}\right)^d}\left(\frac{r_1 - \sigma}{r_2 + \sigma}\right)^d \left(\frac{1 - \left(\frac{r_1 - \sigma}{2\tau}\right)^2}{1 + \left(\frac{2(r_2+\sigma)}{\tau - 2(r_2+\sigma)}\right)^2}\right)^{d/2} \frac{(r_1 - \sigma)^2}{d}.$$

Proof For any unit vector $u \in T_y\mathcal{M}$ we have

$$\begin{aligned}\mathbb{E}\langle u, Z - \mathbb{E}Z\rangle^2 &= \frac{1}{\mathrm{Vol}(Q \cap \mathcal{M}_\sigma)} \int_{Q \cap \mathcal{M}_\sigma} \langle u, Z - \mathbb{E}Z\rangle^2 d\mathrm{Vol}(Z) \\ &\geq \frac{1}{\mathrm{Vol}(B(y, r_2) \cap \mathcal{M}_\sigma)} \int_{B(y,r_1) \cap \mathcal{M}_\sigma} \langle u, (Z - y) - \mathbb{E}(Z - y)\rangle^2 d\mathrm{Vol}(Z)\end{aligned}$$

using the inclusion assumptions, and by adding and subtracting the constant vector y.

We now seek to reduce the domain of integration and perform a change of variables. Since $r_1 \leq \tau/8$, the inverse of the affine projection onto $y + T_y\mathcal{M}$ is injective. Without loss of generality, we assume $y = 0$ and $T_y\mathcal{M}$ is the span of the first d standard orthonormal vectors. Letting f denote the inverse of the affine projection onto $y + T_y\mathcal{M}$, we see that the map

$$\begin{pmatrix} v \\ \beta \end{pmatrix} \longmapsto \begin{pmatrix} v \\ f(v) + \beta \end{pmatrix}$$

is well-defined and injective on $\mathrm{Proj}_{T_y\mathcal{M}}(\mathcal{M} \cap B(y, r_1 - \sigma)) \times (T_y^\perp\mathcal{M} \cap B(0, \sigma))$. Let g denote this map, note that

$$\|x + \beta\| \leq \|x\| + \|\beta\| \leq (r - \sigma) + \sigma = r,$$

for $x \in \mathcal{M} \cap B(y, r_1 - \sigma)$, and hence the image of g is contained in $\mathcal{M}_\sigma \cap B(y, r_1)$. Since the absolute value of the determinant of the Jacobian of g is always 1 (it is lower triangular with ones on the diagonal), employing the change of coordinates in the reduced domain of integration yields

$$\mathbb{E}\langle u, Z - \mathbb{E}Z\rangle^2 \geq \frac{1}{\mathrm{Vol}(B(y, r_2) \cap \mathcal{M}_\sigma)} \int_{\mathcal{A}} \int_{\mathcal{B}} \left\langle \begin{pmatrix} u \\ 0 \end{pmatrix}, \begin{pmatrix} v \\ f(v) + \beta \end{pmatrix} - \mathbb{E}(Z - y) \right\rangle^2 d\beta dv,$$

where

$$\mathcal{A} = \mathrm{Proj}_{T_y\mathcal{M}}(B(y, r_1 - \sigma) \cap \mathcal{M}),\ \mathcal{B} = T_y^{\perp}\mathcal{M} \cap B(0, \sigma).$$

Note that $B(y, \cos(\theta)(r_1 - \sigma)) \cap (y + T_y\mathcal{M}) \subset \mathcal{A}$. Setting $\mathcal{Q} = \mathrm{Proj}_{T_y\mathcal{M}}$, this immediately reduces to

$$\begin{aligned}
\mathbb{E}\langle u, Z - \mathbb{E}Z\rangle^2 &\geq \frac{1}{\mathrm{Vol}(B(y, r_2) \cap \mathcal{M}_\sigma)} \int_{\mathcal{A}} \int_{\mathcal{B}} \langle u, v - \mathbb{E}\mathcal{Q}(Z - y)\rangle^2 d\beta dv \\
&= \frac{\mathrm{Vol}(B_{D-d}(0, \sigma))}{\mathrm{Vol}(B(y, r_2) \cap \mathcal{M}_\sigma)} \int_{\mathcal{A}} \langle u, v - \mathbb{E}\mathcal{Q}(Z - y)\rangle^2 dv \\
&\geq \frac{\mathrm{Vol}(B_{D-d}(0, \sigma))}{\mathrm{Vol}(B(y, r_2) \cap \mathcal{M}_\sigma)} \int_{B_d(0,q)} \langle u, v - \mathbb{E}\mathcal{Q}(Z - y)\rangle^2 dv,
\end{aligned}$$

where $q = \cos(\delta)(r_1 - \sigma)$ and $\delta = \arcsin((r_1 - \sigma)/2\tau)$. Noting that $\int_{B_d(0,q)} \langle u, v\rangle dv = 0$ by symmetry, we now use linearity of the inner product to further reduce the integrand:

$$\begin{aligned}
\mathbb{E}\langle u, Z - \mathbb{E}Z\rangle^2 &\geq \frac{\mathrm{Vol}(B_{D-d}(0, \sigma))}{\mathrm{Vol}(B(y, r_2) \cap \mathcal{M}_\sigma)} \int_{B_d(0,q)} \left(\langle u, v\rangle^2 - 2\langle u, v\rangle\langle u, \mathbb{E}\mathcal{Q}(Z - y)\rangle + \langle u, \mathbb{E}\mathcal{Q}(Z - y)\rangle^2\right) dv \\
&= \frac{\mathrm{Vol}(B_{D-d}(0, \sigma))}{\mathrm{Vol}(B(y, r_2) \cap \mathcal{M}_\sigma)} \int_{B_d(0,q)} \left(\langle u, v\rangle^2 + \langle u, \mathbb{E}\mathcal{Q}(Z - y)\rangle^2\right) dv \\
&\geq \frac{\mathrm{Vol}(B_{D-d}(0, \sigma))}{\mathrm{Vol}(B(y, r_2) \cap \mathcal{M}_\sigma)} \int_{B_d(0,q)} \langle u, v\rangle^2 dv \\
&= \frac{\mathrm{Vol}(B_{D-d}(0, \sigma))\mathrm{Vol}(B_d(0, q))}{\mathrm{Vol}(B(y, r_2) \cap \mathcal{M}_\sigma)} \frac{q^2}{d}.
\end{aligned}$$

By Lemma 16, we then obtain

$$\begin{aligned}
\mathbb{E}\langle u, Z - \mathbb{E}Z\rangle^2 &\geq \left(\left(1 + \frac{\sigma}{\tau}\right)\sqrt{1 + \left(\frac{2(r_2 + \sigma)}{\tau - 2(r_2 + \sigma)}\right)^2}\right)^{-d} \frac{\mathrm{Vol}(B_d(0, q))}{\mathrm{Vol}(B_d(0, r_2 + \sigma))} \frac{q^2}{d} \\
&\geq \frac{1}{4\left(1 + \frac{\sigma}{\tau}\right)^d} \left(\frac{r_1 - \sigma}{r_2 + \sigma}\right)^d \left(\frac{1 - \left(\frac{r_1 - \sigma}{2\tau}\right)^2}{1 + \left(\frac{2(r_2+\sigma)}{\tau - 2(r_2+\sigma)}\right)^2}\right)^{d/2} \frac{(r_1 - \sigma)^2}{d}. \qquad (26)
\end{aligned}$$

Let $V_{d-1}(\Sigma)$ be a subspace corresponding to the first $d - 1$ principal components of Z:

$$V_{d-1} = \operatorname*{argmin}_{\dim(V) = d-1} \mathbb{E}\|Z - \mathbb{E}Z - \mathrm{Proj}_V(Z - \mathbb{E}Z)\|,$$

and note that $\lambda_d(\Sigma) = \max_{0 \neq u \in V_{d-1}^{\perp}} \mathbb{E}\left\langle \frac{u}{\|u\|}, Z - \mathbb{E}Z\right\rangle^2$. Since $\dim(V_{d-1}^{\perp}) = D - d + 1$ and $\dim(T_y\mathcal{M}) = d$, it is easy to see that $V_{d-1}^{\perp} \cap T_y\mathcal{M} \neq \emptyset$. For any $u_* \in V_{d-1}^{\perp} \cap T_y\mathcal{M}$ such that $\|u_*\| = 1$ it follows from Courant-Fischer characterization of $\lambda_d(\Sigma)$ that

$$\lambda_d(\Sigma) \geq \mathbb{E}\left\langle u_*, Z - \mathbb{E}Z\right\rangle^2,$$

and (26) implies the desired bound. ■

The following statement is key to establishing the error bounds for GMRA measured in sup-norm.

Lemma 20 *Assume that conditions of Lemma 19 hold, and let $V_d := V_d(\Sigma)$ be the subspace corresponding to the first d principal components of Z. Then*

$$\sup_{x\in Q} \|x - \mathbb{E}Z - \mathrm{Proj}_{V_d}(x-\mathbb{E}Z)\| \le 2\sigma + \frac{4r_2^2}{\tau} + \frac{r_2}{r_1-\sigma}\sqrt{4\sigma^2 + \frac{16r_2^4}{\tau^2}}\,\gamma(\sigma,\tau,d,r_1,r_2),$$

where $\gamma(\sigma,\tau,d,r_1,r_2) = 4\sqrt{2d}\left(1+\frac{\sigma}{\tau}\right)^{d/2}\left(\frac{r_2+\sigma}{r_1-\sigma}\right)^{d/2}\left(\frac{1+\left(\frac{2(r_2+\sigma)}{\tau-2(r_2+\sigma)}\right)^2}{1-\left(\frac{r_1-\sigma}{2\tau}\right)^2}\right)^{d/4}$.

The proof is given in Appendix 6.2; notice that the term containing $\gamma(\sigma,\tau,d,r_1,r_2)$ is often of smaller order, so that the approximation is essentially controlled by the maximum of σ and $\frac{r_2^2}{\tau}$.

5.2.4 Putting all the bounds together

In this final subsection, we prove Theorem 6. We begin by translating Proposition 3.2 in (Niyogi et al., 2008) into our setting. As before, let $\mathcal{X}_n = \{X_1,\dots,X_n\}$ be an i.i.d. sample from Π, and the ϕ_1 be the constant defined by (10).

Proposition 21 *(Niyogi et al., 2008, Proposition 3.2) Suppose $0<\varepsilon<\frac{\tau}{2}$, and also that n and t satisfy*

$$n \ge \varepsilon^{-d}\frac{1}{\phi_1}\left(\frac{\tau+\sigma}{\tau-\sigma}\right)^d \beta_1\left(\log(\varepsilon^{-d}\beta_2)+t\right), \tag{27}$$

where $\beta_1 = \frac{\mathrm{Vol}_{\mathcal{M}}(\mathcal{M})}{\cos^d(\delta_1)\mathrm{Vol}(B_d(0,1/4))}$, $\beta_2 = \frac{\mathrm{Vol}_{\mathcal{M}}(\mathcal{M})}{\cos^d(\delta_2)\mathrm{Vol}(B_d(0,1/8))}$, $\delta_1 = \arcsin(\varepsilon/8\tau)$, and $\delta_2 = \arcsin(\varepsilon/16\tau)$. Let $\mathcal{E}_{\varepsilon/2,n}$ be the event that

$$\mathcal{Y} = \{Y_j = \mathrm{Proj}_{\mathcal{M}}(X_j)\}_{j=1}^n$$

is $\varepsilon/2$-dense in $\mathcal{M}$ (that is, $\mathcal{M}\subseteq\bigcup_{i=1}^n B(Y_i,\varepsilon/2)$). Then, $\Pi^n(\mathcal{E}_{\varepsilon,n})\ge 1-e^{-t}$, where Π^n is the n-fold product measure of Π.

Proof The proof closely follows the one given in (Niyogi et al., 2008). The only additional observation to make is that, if $\widetilde{\Pi}$ is the pushforward measure of Π under $\mathrm{Proj}_{\mathcal{M}}:\mathcal{M}_\sigma\to\mathcal{M}$, then

$$\begin{aligned}
\widetilde{\Pi}\left(\mathcal{M}\cap B(y,\varepsilon/8)\right) &= \Pi(\mathrm{Proj}_{\mathcal{M}}^{-1}(\mathcal{M}\cap B(y,\varepsilon/8))) \\
&\ge \phi_1 U_{\mathcal{M}_\sigma}(\mathrm{Proj}_{\mathcal{M}}^{-1}(\mathcal{M}\cap B(y,\varepsilon/8))) \\
&= \phi_1 \widetilde{U}_{\mathcal{M}_\sigma}(\mathcal{M}\cap B(y,\varepsilon/8)) \\
&\ge \phi_1\left(\frac{\tau-\sigma}{\tau+\sigma}\right)^d U_{\mathcal{M}}(\mathcal{M}\cap B(y,\varepsilon/8)).
\end{aligned}$$

by Lemma 16. ■

If $\varepsilon \ll \tau$, previous proposition implies that we roughly need $n \geq \mathrm{Const}(\mathcal{M}, d) \left(\frac{1}{\varepsilon}\right)^d \log \frac{1}{\varepsilon}$ points to get an ε-net for $\mathcal{M}$. For the remainder of this section, we identify $\varepsilon := \varepsilon(n,t)$ with the smallest $\varepsilon > 0$ satisfying (27) in the statement of Proposition 21, and we also assume that $\varepsilon < \sigma$. Take $j \in \mathbb{Z}_+$ such that

$$\sigma < 2^{-j-2} < \tau. \tag{28}$$

Let $C_{j,k}$ be the partition of $\mathbb{R}^D$ into Voronoi cells defined by (11). Recall that $T_j = \{a_{j,k}\}_{k=1}^{N(j)} \subset \mathcal{X}_n$ is the set of nodes of the cover tree at level j, and set $z_{j,k} = \mathrm{Proj}_{\mathcal{M}}(a_{j,k})$.

Lemma 22 *With probability $\geq 1 - e^{-t}$, for all j satisfying (28) and $k = 1, \dots, N(j)$,*

$$B\left(z_{j,k}, 2^{-j-2}\right) \subseteq C_{j,k} \text{ and } C_{j,k} \cap \mathcal{M}_\sigma \subseteq B(a_{j,k}, 3 \cdot 2^{-j-2} + 2^{-j+1}) \subseteq B(z_{j,k}, 3 \cdot 2^{-j}). \tag{29}$$

Proof Assume the event $\mathcal{E}_{\varepsilon/2,n} = \{\{Y_1, \dots, Y_n\} \text{ is an } \varepsilon/2 \text{ - net in } \mathcal{M}\}$ occurs. By Proposition 21, $\Pr(\mathcal{E}_{\varepsilon/2,n}) \geq 1 - e^{-t}$.

Since the elements of T_j are 2^{-j}-separated, for any $1 \leq k \leq N(j)$, $B(a_{j,k}, 2^{-j-1}) \subseteq C_{j,k}$. Moreover, since $\sigma \leq 2^{-j-2}$ and $\|a_{j,k} - z_{j,k}\| \leq \sigma$,

$$B(z_{j,k}, 2^{-j-1} - 2^{-j-2}) \subseteq B(z_{j,k}, 2^{-j-1} - \sigma) \subseteq B(a_{j,k}, 2^{-j-1}),$$

hence the inclusion $B\left(z_{j,k}, 2^{-j-2}\right) \subseteq C_{j,k}$ follows.

To show that $C_{j,k} \cap \mathcal{M}_\sigma \subseteq B(a_{j,k}, 3 \cdot 2^{-j-2} + 2^{-j+1})$, pick an arbitrary $z \in \mathcal{M}_\sigma$. Note that on the event $\mathcal{E}_{\varepsilon/2,n}$, there exists $y \in \{Y_1, \dots, Y_n\}$ satisfying $\|z - y\| \leq \varepsilon/2 + \sigma$. Let $x(y) \in \mathcal{X}_n$ be such that $y = \mathrm{Proj}_{\mathcal{M}}(x(y))$. By properties of the cover trees (see Remark 4), there exists $x_* \in T_j$ such that $\|x(y) - x_*\| \leq 2^{-j+1}$. Then

$$\|z - x_*\| \leq \|z - y\| + \|y - x(y)\| + \|x(y) - x_*\| \leq \varepsilon/2 + 2\sigma + 2^{-j+1} \leq 3 \cdot 2^{-j-2} + 2^{-j+1}.$$

Since z was arbitrary, the result follows. Finally, $B(a_{j,k}, 3 \cdot 2^{-j-2} + 2^{-j+1}) \subset B(z_{j,k}, 3 \cdot 2^{-j})$ holds since $\|a_{j,k} - z_{j,k}\| \leq 2^{-j-2}$. ■

We now use Lemma 22 to obtain bounds on the constants θ_i for $i = 1, \dots, 4$ and α. We prove a lemma for each of the assumptions **(A1)**, **(A2)**, and **(A3)** and then collect them as the proof of Theorem 6.

Proof [Proof of Theorem 6] Since the hypotheses of Lemma 22 are satisfied with high probability, we first obtain

$$\begin{aligned}
\Pi(C_{j,k}) &\geq \Pi(B(z_{j,k}, 2^{-j-2})) \\
&\geq \phi_1 U_{\mathcal{M}_\sigma}(B(z_{j,k}, 2^{-j-2})) \\
&= \phi_1 \frac{\mathrm{Vol}(\mathcal{M}_\sigma \cap B(z_{j,k}, 2^{-j-2}))}{\mathrm{Vol}(\mathcal{M}_\sigma)} \\
&\geq \phi_1 \frac{\mathrm{Vol}(\mathrm{Proj}_{\mathcal{M}}^{-1}(\mathcal{M} \cap B(z_{j,k}, 2^{-j-2} - \sigma)))}{\mathrm{Vol}(\mathcal{M}_\sigma)} \\
&\geq \phi_1 \left(\frac{\tau - \sigma}{\tau + \sigma}\right)^d \frac{\cos(\delta)^d \mathrm{Vol}(B_d(0, 2^{-j-2} - \sigma))}{\mathrm{Vol}_{\mathcal{M}}(\mathcal{M})} \\
&\geq \frac{\phi_1 \mathrm{Vol}(B_d(0,1))}{2^{4d}\,\mathrm{Vol}_{\mathcal{M}}(\mathcal{M})} \left(\frac{\tau - \sigma}{\tau + \sigma}\right)^d 2^{-jd}.
\end{aligned}$$

where $\delta = \arcsin((2^{-j-2} - \sigma)/2\tau))$. Thus,

$$\theta_1 \geq \frac{\phi_1 \mathrm{Vol}(B_d(0,1))}{2^{4d}\,\mathrm{Vol}_{\mathcal{M}}(\mathcal{M})} \left(\frac{\tau - \sigma}{\tau + \sigma}\right)^d$$

Since the support is contained in a ball of radius $3 \cdot 2^{-j}$, we easily obtain that $\theta_2 \leq 12$. Finally, it is not difficult to deduce from Lemmas 18 and 19 that

$$\theta_3 \geq \frac{\phi_1/\phi_2}{2^{4d+8}\left(1 + \frac{\sigma}{\tau}\right)^d}, \; \theta_4 \leq \left(2 \vee \frac{2^3 3^4}{\tau^2}\right), \text{ and } \alpha = 1.$$

Lemma 20 together with Lemma 22 imply that

$$\theta_5 \leq \left(2 \vee \frac{4 \cdot 3^2}{\tau}\right) \left(1 + 3 \cdot 2^5 \sqrt{2d} \left(1 + \frac{\sigma}{\tau}\right)^{d/2} \left(\frac{1 + \left(\frac{25}{71}\right)^2}{1 - \frac{1}{9 \cdot 2^{12}}}\right)^{d/4}\right).$$

■

6. Numerical experiments

In this section, we present some numerical experiments consistent with the results above.

6.1 d-dimensional sphere $\mathbb{S}^d$ in $\mathbb{R}^D$.

We consider n points $X_1, \ldots, X_n$ sampled i.i.d. from the uniform distribution on the unit sphere in $\mathbb{R}^{d+1}$

$$\mathbb{S}^d := \{x \in \mathbb{R}^{d+1} : \|x\| = 1\}.$$

We then embed $\mathbb{S}^d$ into $\mathbb{R}^D$ for $D = 100, 1000$ by applying a random orthogonal transformation $\mathbb{R}^{d+1} \to \mathbb{R}^D$. Of course, the actual realization of this projection is irrelevant since

our construction is invariant under orthogonal transformations. After performing this embedding, we add two types of noise. In the first case, we add Gaussian noise $\mathcal{N}(0, \frac{\sigma^2}{D} I_D)$ with mean 0 and covariance matrix $\frac{\sigma^2}{D} I_D$, where the scaling factor $\frac{1}{D}$ is chosen so that $\mathbb{E}\|\eta\|^2 = \sigma^2$, and in fact $\|\eta\|^2$ is highly concentrated around its mean σ^2. In this way (up to a small number of samples for which $||\eta_i||^2 \gg \sigma^2$), this data set almost satisfies the $(1, (1 + \frac{1}{\sqrt{D}})\sigma)$-model assumption. We consider the behavior of the $L^2(\Pi)$ error squared (MSE) in these case in Figure 1, and the rate of approximation at the optimal scale, as the number of samples varies in Figure 3, where it is compared to the rates obtained in Corollary 7. From Figure 1, we see that the approximations obtained satisfy our bound, and are typically better even for a modest number of samples in dimensions non-trivially low (e.g. 8000 samples on $\mathbb{S}^8$). In fact, the robustness with respect to sampling is such that the plots barely change from row to row.

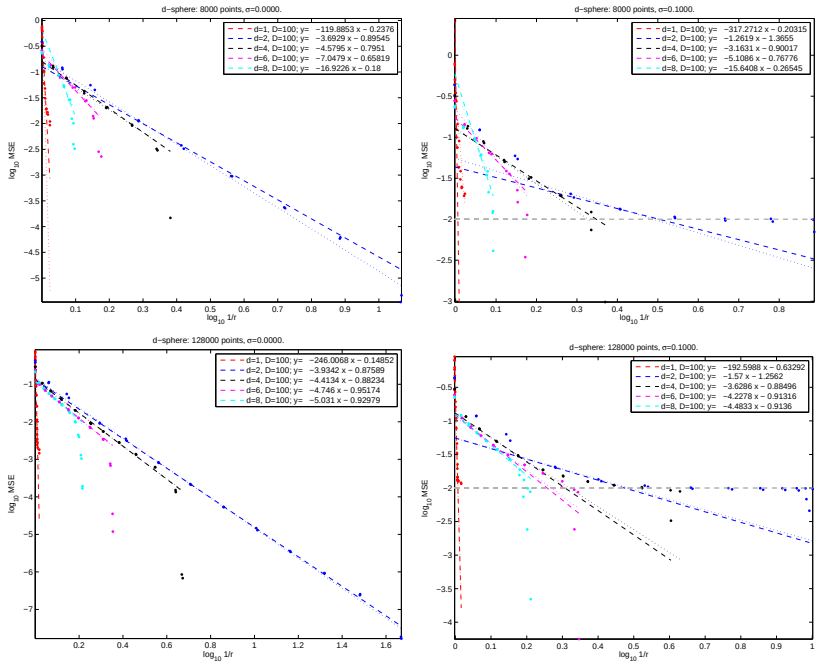


Figure 1: Experiment with $\mathbb{S}^d$, in the noiseless case (left column) and Gaussian noise case (right column). The rows correspond to different values of $n \in \{8000, 128000\}$. In the plots the dots represent the $L^2(\Pi)$ error squared (or MSE) of GMRA approximations (as in (5)) as a function of scale j; more precisely the abscissa is in terms of $\log_{10}(1/r_j)$, where r_j is the mean radius of $C_{j,k}$ for a fixed j, and the ordinate is $\log_{10} \mathrm{MSE}_j$, where MSE_j is the mean squared error of the GMRA approximation at scale j. Different colors correspond to different intrinsic dimensions d. The two cases $D = 100, 1000$ use the same colors for both the dots and the lines, all of which are essentially superimposed since our results are independent of the ambient dimension D. For each dimension we fit a line to measure the decay, which should be $O(r^{-4})$ independently of d. On the right we have the noiseless case. The horizontal dashed gray line represents the noise level σ^2; the approximation error flattens out at that level.

The second type of noise is uniform in the radial direction, i.e. we let $\eta \sim \mathrm{Unif}[1-\sigma, 1+\sigma]$ and each noisy point is generated by $\tilde{X}_i = X_i + \eta_i \frac{X_i}{||X_i||}$. This is an example where the noise is not independent of X, but yet our $(1,\sigma)$-model assumptions are satisfied. The results are summarized in Figure 2, with the rate of approximation at the optimal scale again in Figure 3.

We considered various settings of the parameters, namely all combinations of: $d \in \{1,2,4,6,8\}$, $n \in \{8000, 16000, 32000, 64000, 128000\}$, $D \in \{100, 1000\}$, $\sigma \in \{0, 0.05, 0.1\}$. We only display some of the results for reasons of space[2].

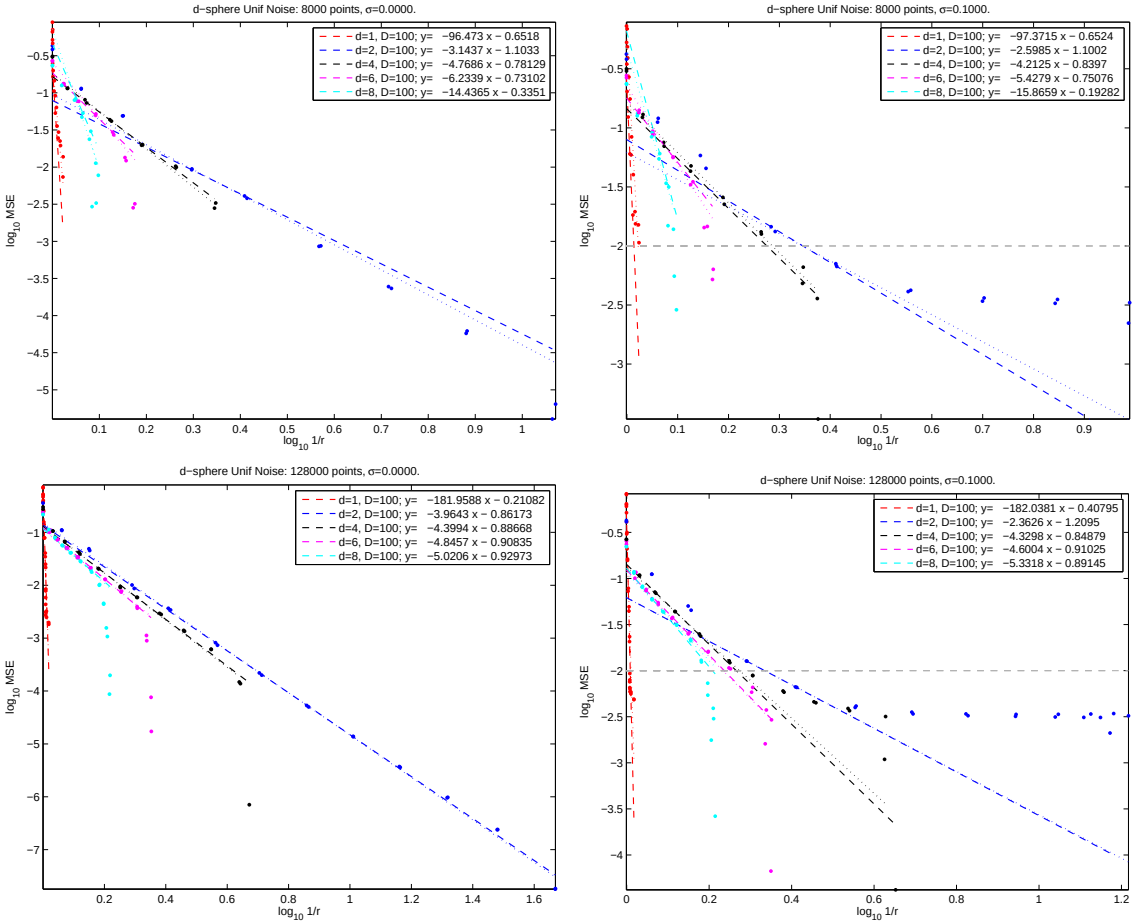


Figure 2: MSE as a function of scale r for $\mathbb{S}^d$, for different values of $n =$, d and σ, the width of uniform noise in the radial direction. Note that the variance of the noise is $\sigma^2/3$, which accounts for the fact that the MSE, in the noisy case (see insets in the right column), approaches a level slightly lower than σ^2.

2. The code provided at www.math.duke.edu/~mauro/code.html can generate all the figures, re-create the data sets, and is easily modified to do more experiments.

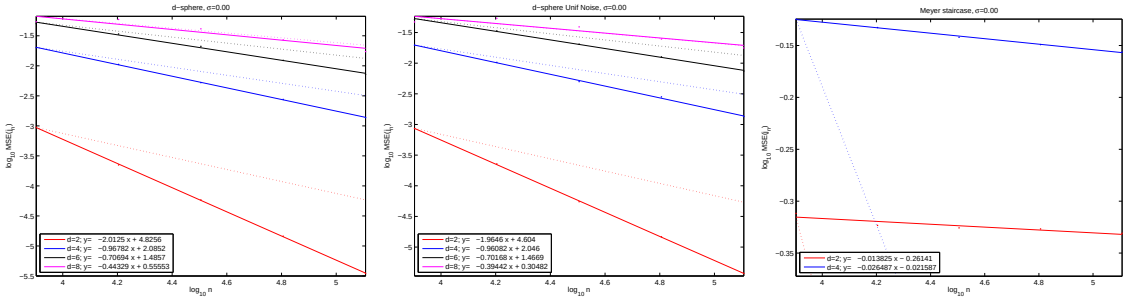


Figure 3: For the three examples considered in this section (left: $\mathbb{S}^d$; center: $\mathbb{S}^d$; right: Meyer's staircase), in the noiseless case, we consider the MSE error, i.e. $L^2(\Pi)$ squared error (as defined in (5)) at the optimal scale j_n (as in the proof of Corollary 7) as a function of the number of points n for $\mathbb{S}^d$ with Gaussian noise (let) and "uniform radial" noise as described in the text. This is compared with the rates predicted by Corollary 7. We attribute the slightly better performance of GMRA in dimension $d \geq 2$, as compared to the predicted rate, to concentration phenomena on the sphere, as discussed in Little (2012). In the case of the Meyer staircase, the predicted rates are far from the ones measured experimentally: we believe this is due to small reach of the manifold that affects the constants in front of the decay rate. Only a very large number of samples (even larger than the $128,000$ used here) would be fine enough to reveal the optimal rate of decay.

6.2 Meyer staircase

We consider the (d-dimensional generalization of) Y. Meyer's staircase. Consider the unit cube $Q = [-\frac{1}{2}, \frac{1}{2}]^d$ and a Gaussian $\mathcal{N}(x_0, \delta^2 I_D)$ restricted to Q. As x_0 varies in Q this describes a smooth D-dimensional manifold in $L^2(Q)$. This example may be discretized, in particular a grid Γ_D of D points (obtained by subdividing in $D^{-\frac{1}{d}}$ parts along each dimension) in Q may be generated. n points may be sampled from this manifold by sampling $x_1, \ldots, x_n$ uniformly at random in Q, obtaining a set $\{\mathcal{N}(x_i, \delta^2 I_D)|_{\Gamma_D}\}_{i=1,\ldots,n}$ of n points in $\mathbb{R}^D$. This is what we call a sample from the Meyer staircase. This example is not artificial: for example a set of $2 - D$ images obtained by taking a white shape on a black background and translating the shape around has many similarities with the Meyer staircase.

The manifold associated with the Meyer staircase is poorly approximated by subspaces of dimension smaller than $O(D \wedge 1/\delta^D)$, and besides spanning many dimensions in $\mathbb{R}^D$, it has a small reach, depending on d, D, δ. In our examples we considered $n = 8000, 16000, 32000, 640000, 128000$, $d = 1, 2, 4$, $D = 2^9$, and $\delta = \frac{100}{8000^{\frac{1}{d}}}$. We consider the noiseless case, as well as the case where Gaussian noise $\mathcal{N}(0, \frac{1}{D} I_D)$ is added to the data.

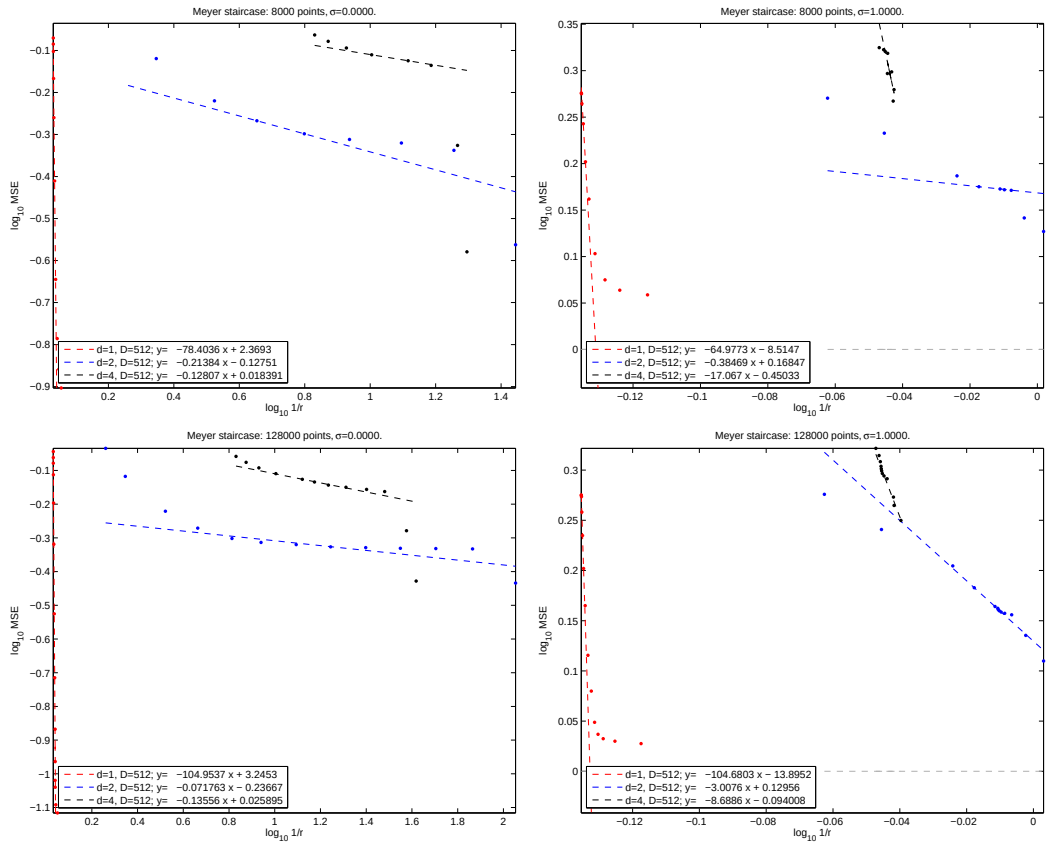


Figure 4: MSE as a function of scale r for the d-dimensional Meyer staircase, for different values of $n =$, d and σ, standard deviation of Gaussian noise $\mathcal{N}(0, \frac{\sigma^2}{D})$. The small reach of Meyer's staircase make it harder to approximate, and make it much more susceptible to noise.

Acknowledgments

The authors gratefully acknowledge support from NSF DMS-0847388, NSF DMS-1045153, ATD-1222567, CCF-0808847, AFOSR FA9550-14-1-0033, DARPA N66001-11-1-4002. We would also like to thank Mark Iwen for his insightful comments.

Appendix A: proof of Proposition 8

For the first inequality, let

$$A = \begin{pmatrix} I \\ X \end{pmatrix} \text{ and } B = \begin{pmatrix} Y \\ 0 \end{pmatrix},$$

and for every $T \subset [d]$, we let V_T denote the volume of $\{a_i\}_{i \in T^c} \cup \{b_i\}_{i \in T}$, where a_i and b_i denote the ith columns of A and B respectively. By submultilinearity of the volume we have

$$\mathrm{Vol}(A+B) \leq \sum_{T \in 2^{[d]}} V_T,$$

where $2^{[d]} = \{S : S \subset \{1, \dots, d\}\}$. We now show that $V_T \leq q^{|T|}\mathrm{Vol}(A)$ for every $T \in 2^{[d]}$. The bound $\|Y\| \leq q$ implies $\|y_i\| \leq q$ for all $i = 1, \dots, d$, and so the fact that the volume is a submultiplicative function implies that

$$V_T \leq q^{|T|}\mathrm{Vol}(A_{T^c}).$$

On the other hand, letting $a_1^\perp$ be the orthogonal projection of a_1 onto $\mathrm{span}^\perp\{a_i\}_{i=2}^d$, we note that $\|a_1^\perp\| \geq 1$, and thus

$$\mathrm{Vol}(A_{\{1\}^c}) \leq \|a_1^\perp\|\mathrm{Vol}(A_{\{1\}^c}) = \mathrm{Vol}(A).$$

By induction and invariance of the volume under permutations, we see that $\mathrm{Vol}(A_{T^c}) \leq \mathrm{Vol}(A)$ for all $T \in 2^{[d]}$. Thus,

$$\mathrm{Vol}(A+B) \leq \sum_{T \in 2^{[d]}} q^{|T|}\mathrm{Vol}(A) = (1+q)^d \mathrm{Vol}(A).$$

For the second inequality, since Y is symmetric, we can represent it as $Y = F - G$ where F and G are symmetric positive semidefinite, $FG = GF = 0$, and $\|F\|, \|G\| \leq \|Y\|$. Indeed, if $Y = Q\Lambda Q^T$ is the eigenvalue decomposition of Y with $\Lambda = \mathrm{diag}(\lambda)$, set $\lambda_+ := (\max(0, \lambda_1), \dots, \max(0, \lambda_d))^T$, $\lambda_- := \lambda_+ - \lambda$, and define $F := Q\,\mathrm{diag}(\lambda_+)Q^T$, $G = Q\,\mathrm{diag}(\lambda_-)Q^T$.

Recall the *matrix determinant lemma*: let $T \in \mathbb{R}^{k \times k}$ be invertible, and let $U, V \in \mathbb{R}^{k \times l}$. Then

$$\mathrm{Vol}(T + UV^T) = \mathrm{Vol}(I + V^T T^{-1} U)\mathrm{Vol}(T).$$

Applying it in our case with $U = \begin{pmatrix} \sqrt{F} - \sqrt{G} \\ 0 \end{pmatrix}$, $V = \begin{pmatrix} \sqrt{F} + \sqrt{G} \\ 0 \end{pmatrix}$, and $T = \begin{pmatrix} I & X^T \\ X & -I \end{pmatrix}$, we have that

$$\mathrm{Vol}\begin{pmatrix} I+Y & X^T \\ X & -I \end{pmatrix} = \mathrm{Vol}\left(I + \begin{pmatrix} \sqrt{F} + \sqrt{G} \\ 0 \end{pmatrix}^T \begin{pmatrix} I & X^T \\ X & -I \end{pmatrix}^{-1} \begin{pmatrix} \sqrt{F} - \sqrt{G} \\ 0 \end{pmatrix} \right) \mathrm{Vol}\begin{pmatrix} I & X^T \\ X & -I \end{pmatrix}.$$

By orthogonality of the columns in

$$\begin{pmatrix} I \\ X \end{pmatrix}$$

with the columns in

$$\begin{pmatrix} X^T \\ -I \end{pmatrix},$$

we have that

$$\left\| \begin{pmatrix} I & X^T \\ X & -I \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} \right\| \geq \left\| \begin{pmatrix} u \\ v \end{pmatrix} \right\|,$$

and hence

$$\left\| \begin{pmatrix} \sqrt{F} + \sqrt{G} \\ 0 \end{pmatrix}^T \begin{pmatrix} I & X^T \\ X & -I \end{pmatrix}^{-1} \begin{pmatrix} \sqrt{F} - \sqrt{G} \\ 0 \end{pmatrix} \right\| \leq \sqrt{q} \cdot 1 \cdot \sqrt{q} = q.$$

Therefore, we conclude that

$$\mathrm{Vol}\left(I + \begin{pmatrix} \sqrt{F} + \sqrt{G} \\ 0 \end{pmatrix}^T \begin{pmatrix} I & X^T \\ X & -I \end{pmatrix}^{-1} \begin{pmatrix} \sqrt{F} - \sqrt{G} \\ 0 \end{pmatrix} \right) \geq (1-q)^d,$$

and combining this with the expression from the matrix determinant lemma completes the proof.

Appendix B: angle between the tangent and principal component subspaces

Let $Q \subset \mathbb{R}^D$ be such that $B(y, r_1) \subset Q$ and $\mathcal{M}_\sigma \cap Q \subset B(y, r_2)$ for some $y \in \mathcal{M}$ and $\sigma < r_1 < r_2 < \tau/8 - \sigma$. Assume that Z is drawn from $U_{\mathcal{M}_\sigma \cap Q}$, let Σ be the covariance matrix of Z and $V_d := V_d(\Sigma)$ - the subspace corresponding to the first d principal components of Z.

Let $\alpha \in [0,1]$ be such that $\cos(\phi) := \min_{u \in V_d, \|u\|=1} \max_{v \in T_y\mathcal{M}, \|v\|=1} |\langle u, v \rangle| = \sqrt{1-\alpha^2}$ is the cosine of the angle between $T_y\mathcal{M}$ and V_d. Then there exists a unit vector $u_* \in (V_d)^\perp$ such that

$$\max_{v \in T_y\mathcal{M}, \|v\|=1} |\langle u_*, v \rangle| \geq \alpha.$$

Indeed, let $u' \in V_d$, $v' \in T_y\mathcal{M}$ be unit vectors such that $\cos(\phi) = \langle u', v' \rangle$, Note that $\sqrt{1-\alpha^2}$ is equal to the smallest absolute value among the nonzero singular values of the operator $\mathrm{Proj}_{T_y\mathcal{M}} \mathrm{Proj}_{V_d}$. Since the spectra of the operators $\mathrm{Proj}_{T_y\mathcal{M}} \mathrm{Proj}_{V_d}$ and $\mathrm{Proj}_{V_d} \mathrm{Proj}_{T_y\mathcal{M}}$ coincide by the well-known fact from linear algebra, we have that

$$\min_{u \in V_d, \|u\|=1} \max_{v \in T_y\mathcal{M}, \|v\|=1} |\langle u, v \rangle| = \min_{v \in T_y\mathcal{M}, \|v\|=1} \max_{u \in V_d, \|u\|=1} |\langle u, v \rangle|.$$

In other words, $\mathrm{Proj}_{T_y\mathcal{M}}(u') = \langle u', v' \rangle v'$ and $\mathrm{Proj}_{V_d}(v') = \langle u', v' \rangle u'$. This implies that there exists a unit vector $u_* \in (V_d)^\perp$ such that $v' = \langle v', u' \rangle u' + \langle v', u_* \rangle u_*$, hence $\langle u_*, v' \rangle^2 = 1 - \langle v', u' \rangle^2 = \alpha^2$, so u_* satisfies the requirement.

To simplify the expressions, let

$$\zeta = \frac{1}{\mathrm{Vol}(Q \cap \mathcal{M}_\sigma)}.$$

We shall now construct upper and lower bounds for

$$\zeta \int_{Q\cap\mathcal{M}_\sigma} \langle u_*, x-\mathbb{E}Z - \mathrm{Proj}_{V_d}(x-\mathbb{E}Z)\rangle^2 \, d\mathrm{Vol}(x) = \zeta \int_{Q\cap\mathcal{M}_\sigma} \langle u_*, x-\mathbb{E}Z\rangle^2 \, d\mathrm{Vol}(x)$$

which together yield an estimate for α. Write $u_* = u_*^{||} + u_*^{\perp}$, where $u_*^{||} \in T_y\mathcal{M}$ and $u_*^{\perp} \in T_y^{\perp}\mathcal{M}$. By our choice of u_*, we clearly have that $\|u_*^{||}\| = \max_{v\in T_y\mathcal{M}, \|v\|=1} \langle u_*, v\rangle \geq \alpha$. Using the elementary inequality $(a+b)^2 \geq \frac{a^2}{2} - b^2$, we further deduce that

$$\begin{aligned}\zeta \int_{Q\cap\mathcal{M}_\sigma} \langle u_*, x-\mathbb{E}Z\rangle^2 \, d\mathrm{Vol}(x) \geq \zeta \int_{Q\cap\mathcal{M}_\sigma} \frac{1}{2}\left\langle u_*^{||}, x-\mathbb{E}Z\right\rangle^2 d\mathrm{Vol}(x) \\ - \zeta \int_{Q\cap\mathcal{M}_\sigma} \left\langle u_*^{\perp}, x-\mathbb{E}Z\right\rangle^2 d\mathrm{Vol}(x).\end{aligned} \tag{30}$$

It follows from the proof of Lemma 19 that

$$\zeta \int_{Q\cap\mathcal{M}_\sigma} \frac{1}{2}\left\langle u_*^{||}, x-\mathbb{E}Z\right\rangle^2 d\mathrm{Vol}(x) \geq \frac{\alpha^2}{8\left(1+\frac{\sigma}{\tau}\right)^d}\left(\frac{r_1-\sigma}{r_2+\sigma}\right)^d \left(\frac{1-\left(\frac{r_1-\sigma}{2\tau}\right)^2}{1+\left(\frac{2(r_2+\sigma)}{\tau-2(r_2+\sigma)}\right)^2}\right)^{d/2} \frac{(r_1-\sigma)^2}{d}.$$

For the last term in (30), Lemma 18 (see equation (25)) gives

$$\begin{aligned}\zeta \int_{Q\cap\mathcal{M}_\sigma} \left\langle u_*^{\perp}, x-\mathbb{E}Z\right\rangle^2 d\mathrm{Vol}(x) &\leq \zeta \int_{Q\cap\mathcal{M}_\sigma} \|x-\mathbb{E}Z - \mathrm{Proj}_{T_y\mathcal{M}}(x-\mathbb{E}Z)\|^2 d\mathrm{Vol}(x) \\ &\leq 2\sigma^2 + \frac{8r_2^4}{\tau^2},\end{aligned}$$

hence (30) yields

$$\begin{aligned}\zeta \int_{Q\cap\mathcal{M}_\sigma} \langle u_*, x-\mathbb{E}Z\rangle^2 \, d\mathrm{Vol}(x) \geq &\frac{\alpha^2}{8\left(1+\frac{\sigma}{\tau}\right)^d}\left(\frac{r_1-\sigma}{r_2+\sigma}\right)^d \left(\frac{1-\left(\frac{r_1-\sigma}{2\tau}\right)^2}{1+\left(\frac{2(r_2+\sigma)}{\tau-2(r_2+\sigma)}\right)^2}\right)^{d/2} \frac{(r_1-\sigma)^2}{d} \\ &- 2\sigma^2 - \frac{8r_2^4}{\tau^2}.\end{aligned} \tag{31}$$

On the other hand, invoking (25) once again, we have

$$\zeta \int_{Q\cap\mathcal{M}_\sigma} \langle u_*, x-\mathbb{E}Z\rangle^2 \, d\mathrm{Vol}(x) \leq 2\sigma^2 + \frac{8r_2^4}{\tau^2}.$$

Combined with (31), this gives

$$\frac{\alpha^2}{8\left(1+\frac{\sigma}{\tau}\right)^d}\left(\frac{r_1-\sigma}{r_2+\sigma}\right)^d \left(\frac{1-\left(\frac{r_1-\sigma}{2\tau}\right)^2}{1+\left(\frac{2(r_2+\sigma)}{\tau-2(r_2+\sigma)}\right)^2}\right)^{d/2} \frac{(r_1-\sigma)^2}{d} \leq 4\sigma^2 + \frac{16r_2^4}{\tau^2}, \tag{32}$$

and the upper bound for α follows. We are ready to prove Lemma 20.

Proof of Lemma 20

Notice that for any $x \in Q \cap \mathcal{M}_\sigma$,

$$x - \mathbb{E}Z - \operatorname{Proj}_{V_d}(x - \mathbb{E}Z) = x - y - \operatorname{Proj}_{T_y\mathcal{M}}(x-y) + \underbrace{y - \mathbb{E}Z - \operatorname{Proj}_{T_y\mathcal{M}}(y - \mathbb{E}Z)}_{\operatorname{Proj}_{(T_y\mathcal{M})^\perp}(y-\mathbb{E}Z)} + (\operatorname{Proj}_{T_y\mathcal{M}} - \operatorname{Proj}_{V_d})(x - \mathbb{E}Z). \tag{33}$$

It follows from (24) that

$$\|x - y - \operatorname{Proj}_{T_y\mathcal{M}}(x-y)\| = \left\|\operatorname{Proj}_{T_y^\perp\mathcal{M}}(x-y)\right\| \le \sigma + \frac{2r_2^2}{\tau}.$$

Next,

$$\begin{aligned}\|\operatorname{Proj}_{(T_y\mathcal{M})^\perp}(y - \mathbb{E}Z)\| &= \frac{1}{\operatorname{Vol}(Q\cap\mathcal{M}_\sigma)}\left\|\int_{Q\cap\mathcal{M}_\sigma}\operatorname{Proj}_{T_y^\perp\mathcal{M}}(y-z)d\operatorname{Vol}(z)\right\| \\ &\le \frac{1}{\operatorname{Vol}(Q\cap\mathcal{M}_\sigma)}\int_{Q\cap\mathcal{M}_\sigma}\left\|\operatorname{Proj}_{T_y^\perp\mathcal{M}}(z-y)\right\| d\operatorname{Vol}(z) \\ &\le \sigma + \frac{2r_2^2}{\tau}.\end{aligned}$$

Finally, it is easy to see that

$$\begin{aligned}\|(\operatorname{Proj}_{T_y\mathcal{M}} - \operatorname{Proj}_{V_d})(x-\mathbb{E}Z)\| \le{}& \|\operatorname{Proj}_{T_y\mathcal{M}}(x-\mathbb{E}Z) - \operatorname{Proj}_{V_d}\operatorname{Proj}_{T_y\mathcal{M}}(x-\mathbb{E}Z)\| \\ &+ \|\operatorname{Proj}_{T_y^\perp\mathcal{M}}(x-y)\| + \|\operatorname{Proj}_{T_y^\perp\mathcal{M}}(\mathbb{E}Z - y)\|.\end{aligned}$$

Let $u_x := \frac{\operatorname{Proj}_{T_y\mathcal{M}}(x-\mathbb{E}Z)}{\|\operatorname{Proj}_{T_y\mathcal{M}}(x-\mathbb{E}Z)\|}$ and note that for any $x \in Q\cap\mathcal{M}_\sigma$, $\|\operatorname{Proj}_{T_y\mathcal{M}}(x-\mathbb{E}Z)\| \le 2r_2$, hence

$$\begin{aligned}\|\operatorname{Proj}_{T_y\mathcal{M}}(x-\mathbb{E}Z) - \operatorname{Proj}_{V_d}\operatorname{Proj}_{T_y\mathcal{M}}(x-\mathbb{E}Z)\|^2 &\le (2r_2)^2\left(1 - \|\operatorname{Proj}_{V_d}u_x\|^2\right) \\ &\le 4r_2^2\left(1 - \min_{u\in T_y\mathcal{M},\|u\|=1}\max_{v\in V_d,\|v\|=1}\langle u,v\rangle^2\right) \\ &= 4r_2^2\alpha^2.\end{aligned}$$

Combining the previous bounds with (32) and (33), we obtain the result.

References

M. Aharon, M. Elad, and A. Bruckstein. K-SVD: Design of dictionaries for sparse representation. volume 5, pages 9–12, 2005.

W. K. Allard, G. Chen, and M. Maggioni. Multi-scale geometric methods for data sets II: Geometric multi-resolution analysis. *Applied and Computational Harmonic Analysis*, 32 (3):435–462, 2012.

M. Belkin and P. Niyogi. Using manifold structure for partially labelled classification. *Advances in NIPS*, 15, 2003.

A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor. In *Proceedings of ICML*, pages 97–104. ACM, 2006.

P. S. Bradley and O. L. Mangasarian. K-plane clustering. *Journal of Global Optimization*, 16(1):23–32, 2000.

F. Camastra and A. Vinciarelli. Intrinsic dimension estimation of data: an approach based on Grassberger–Procaccia's algorithm. *Neural Processing Letters*, 14(1):27–34, 2001.

G. Canas, T. Poggio, and L. Rosasco. Learning manifolds with k-means and k-flats. In *Advances in NIPS*, 25, pages 2474–2482, 2012.

E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, (6):2313–2351, 2007. math.ST/0506081.

E. Causevic, R.R. Coifman, R. Isenhart, A. Jacquin, E.R. John, M. Maggioni, L.S. Prichep, and F.J. Warner. QEEG-based classification with wavelet packets and microstate features for triage applications in the ER. Proceedings of ICASSP, volume 3., May 2006.

G. Chen and M. Maggioni. Multiscale geometric and spectral analysis of plane arrangements. In *Proceedings of CVPR*, 2011.

G. Chen and M.Maggioni. Multiscale geometric wavelets for the analysis of point clouds. *Proceedings of CISS*, pages 1–6, 2010.

G. Chen, A.V. Little, and M. Maggioni. Multi-resolution geometric analysis for data in high dimensions. *Excursions in Harmonic Analysis,* pages 259–285, 2013.

G. Chen, A.V. Little, M. Maggioni, and L. Rosasco. *Wavelets and Multiscale Analysis: Theory and Applications*. Springer Verlag, 2011b.

G. Chen and G. Lerman. Foundations of a multi-way spectral clustering framework for hybrid linear modeling. *Foundations of Computational Mathematics*, 9(5):517–558, 2009.

G. Chen, M. Iwen, S. Chin, and M. Maggioni. A fast multiscale framework for data in high-dimensions: Measure estimation, anomaly detection, and compressive measurements. In *Visual Communications and Image Processing (VCIP), 2012*, pages 1–6, 2012.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998.

P. Ciaccia, M. Patella, F. Rabitti, and P. Zezula. Indexing metric spaces with m-tree. In *SEBD*, volume 97, pages 67–86, 1997.

R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *PNAS*, 102(21):7426–7431, 2005a. doi: 10.1073/pnas.0500334102.

R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: multiscale methods. *PNAS*, 102(21):7432–7438, 2005b. doi: 10.1073/pnas.0500334102.

R.R. Coifman and M. Maggioni. Diffusion wavelets. *Appl. Comp. Harm. Anal.*, 21(1):53–94, July 2006.

R.R. Coifman, S. Lafon, M. Maggioni, Y. Keller, A.D. Szlam, F.J. Warner, and S.W. Zucker. Geometries of sensor outputs, inference, and information processing. Defense and Security Symposium, 2006.

G. David and S. Semmes. *Analysis of and on uniformly rectifiable sets*, volume 38 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1993.

C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

D. Donoho. Compressed sensing. *IEEE Tran. on Information Theory*, 52(4):1289–1306, April 2006.

D. L. Donoho. Wedgelets: nearly-minimax estimation of edges. *Ann. Statist*, pages 859–897, 1999.

D. L. Donoho and C. Grimes. Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10): 5591–5596, 2003.

D. L. Donoho and C. Grimes. When does Isomap recover the natural parameterization of families of articulated images? Technical report, 2002.

A. Eftekhari and M. B. Wakin. New analysis of manifold embeddings and signal recovery from compressive measurements. *arXiv:1306.4748*, 2013.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *Proceedings of CVPR 2009*, pages 2790–2797. IEEE, 2009.

H. Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93 (3):418–491, 1959.

M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax manifold estimation. *J. Mach. Learn. Res.*, 13(1):1263–1291, May 2012a.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Manifold estimation and singular deconvolution under hausdorff loss. *The Annals of Statistics*, 40(2):941–963, 2012b.

C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *The Annals of Statistics*, 40(2):941–963, 2012c.

A. Gray. *Tubes*, volume 221 of *Progress in Mathematics*. Birkhäuser Verlag, Basel, second edition, 2004.

R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert. Sample complexity of dictionary learning and other matrix factorizations. *arXiv:1312.3790*, 2013.

J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman. Clustering appearances of objects under varying illumination conditions. In *Proceedings of CVPR 2003*, volume 1, pages I–11. IEEE, 2003.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(4):17–441,498–520, 1933.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 27:321–77, 1936.

M. A. Iwen and M. Maggioni. Approximation of points on low-dimensional manifolds via random linear projections. *Inference & Information*, 2(1):1–31, 2013.

P. W. Jones. Rectifiable sets and the traveling salesman problem. *Inventiones Mathematicae*, 102(1):1–15, 1990.

P.W. Jones, M. Maggioni, and R. Schul. Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels. *Proc. Nat. Acad. Sci.*, 105(6):1803–1808, Feb. 2008.

P.W. Jones, M. Maggioni, and R. Schul. Universal local manifold parametrizations via heat kernels and eigenfunctions of the Laplacian. *Ann. Acad. Scient. Fen.*, 35:1–44, January 2010.

D. R. Karger and M. Ruhl. Finding nearest neighbors in growth-restricted metrics. In *Proceedings of the 34th annual ACM symposium on Theory of computing*, pages 741–750. ACM, 2002.

V. I. Koltchinskii. Empirical geometry of multivariate data: a deconvolution approach. *The Annals of Statistics*, pages 591–629, 2000.

K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Comput.*, 15(2):349–396, February 2003.

E. Levina and P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in NIPS*, pages 777–784, 2004.

M. S. Lewicki, T. J. Sejnowski, and H. Hughes. Learning overcomplete representations. *Neural Computation*, 12:337–365, 1998.

A.V. Little, Y.-M. Jung, and M. Maggioni. Multiscale estimation of intrinsic dimensionality of data sets. In *Proc. A.A.A.I.*, 2009.

M. Maggioni, L. Rosasco and A. V. Little. Multiscale geometric methods for data sets I: Multiscale SVD, noise and curvature. Technical report, MIT-CSAIL-TR-2012-029/CBCL-310, MIT, Cambridge, MA, September 2012. URL http://dspace.mit.edu/handle/1721.1/72597.

G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of ICML*, pages 663–670, 2010.

Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(9):1546–1562, 2007.

Y. Ma, A. Y. Yang, H. Derksen, and R. Fossum. Estimation of subspace arrangements with applications in modeling and segmenting mixed data. *SIAM review*, 50(3):413–458, 2008.

A. Maurer and M. Pontil. K -dimensional coding schemes in hilbert spaces. *Information Theory, IEEE Transactions on*, 56(11):5839–5846, Nov 2010a.

A. Maurer and M. Pontil. K-dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010b.

S. Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *arXiv:1112.5448*, 2013.

P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry*, 39:419–441, 2008.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, (37), 1997.

K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

G. Peyré. Sparse modeling of textures. *Journal of Mathematical Imaging and Vision*, 34 (1):17–31, 2009.

M. Protter and M. Elad. Sparse and redundant representations and motion-estimation-free algorithm for video denoising. *Optical Engineering+ Applications*, 2007.

I. U. Rahman, I. Drori, V. C. Stodden, D. L. Donoho, and P. Schröder. Multiscale representations for manifold-valued data. *Multiscale Modeling & Simulation*, 4(4):1201–1232, 2005.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Y. Sugaya and K. Kanatani. Multi-stage unsupervised learning for multi-body motion segmentation. *IEICE TRANSACTIONS on Information and Systems*, 87(7):1935–1942, 2004.

J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999.

D. Vainsencher, S. Mannor, and A. M. Bruckstein. The sample complexity of dictionary learning. *J. Mach. Learn. Res.*, 12:3259–3281, November 2011.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(12):1945–1959, 2005.

M. B. Wakin, D. L. Donoho, H. Choi, and R. G. Baraniuk. The multiscale structure of non-differentiable image manifolds. In *SPIE Wavelets XI*, pages 59141B–59141B. International Society for Optics and Photonics, 2005.

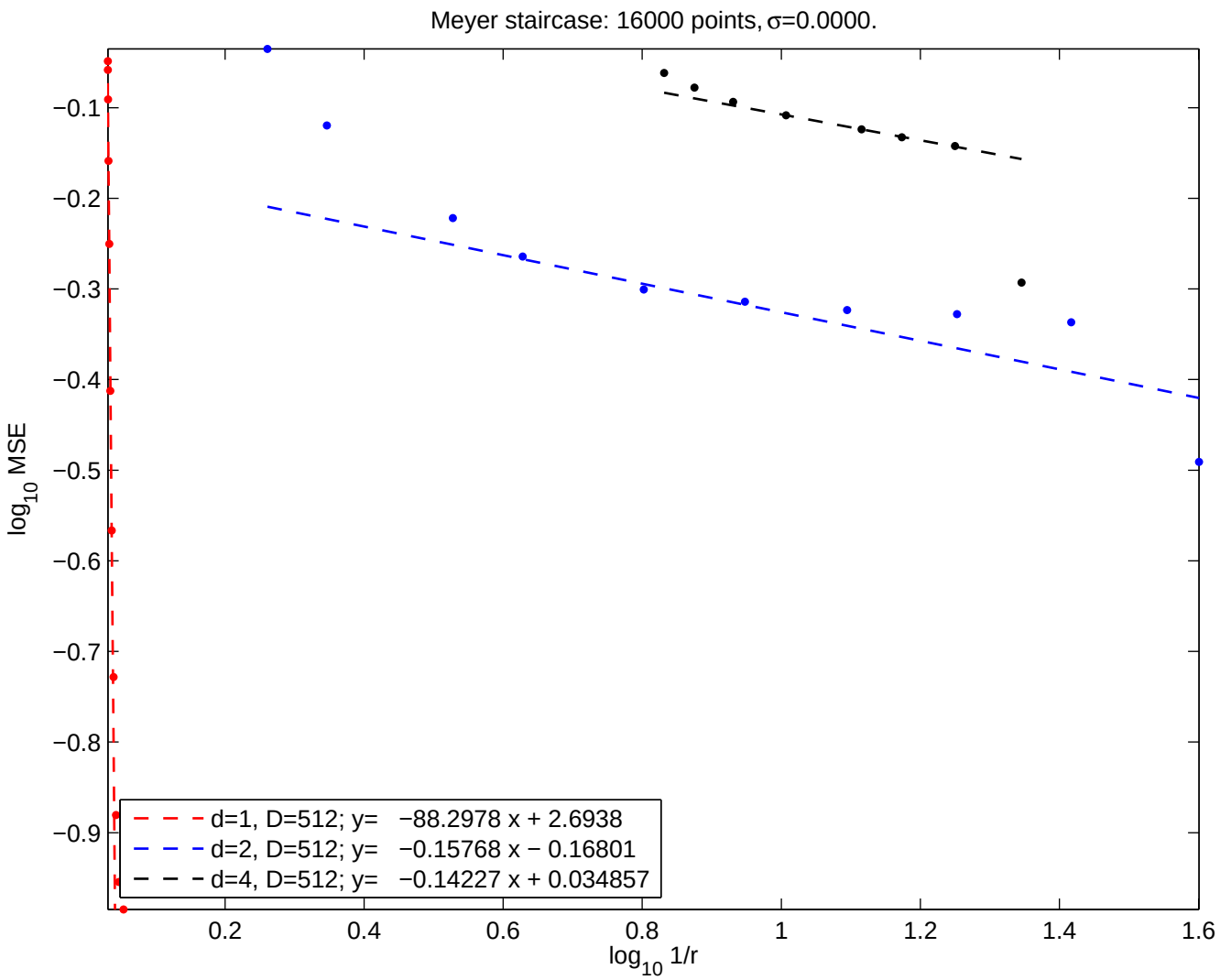
J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *Computer Vision–ECCV 2006*, pages 94–106. Springer, 2006.

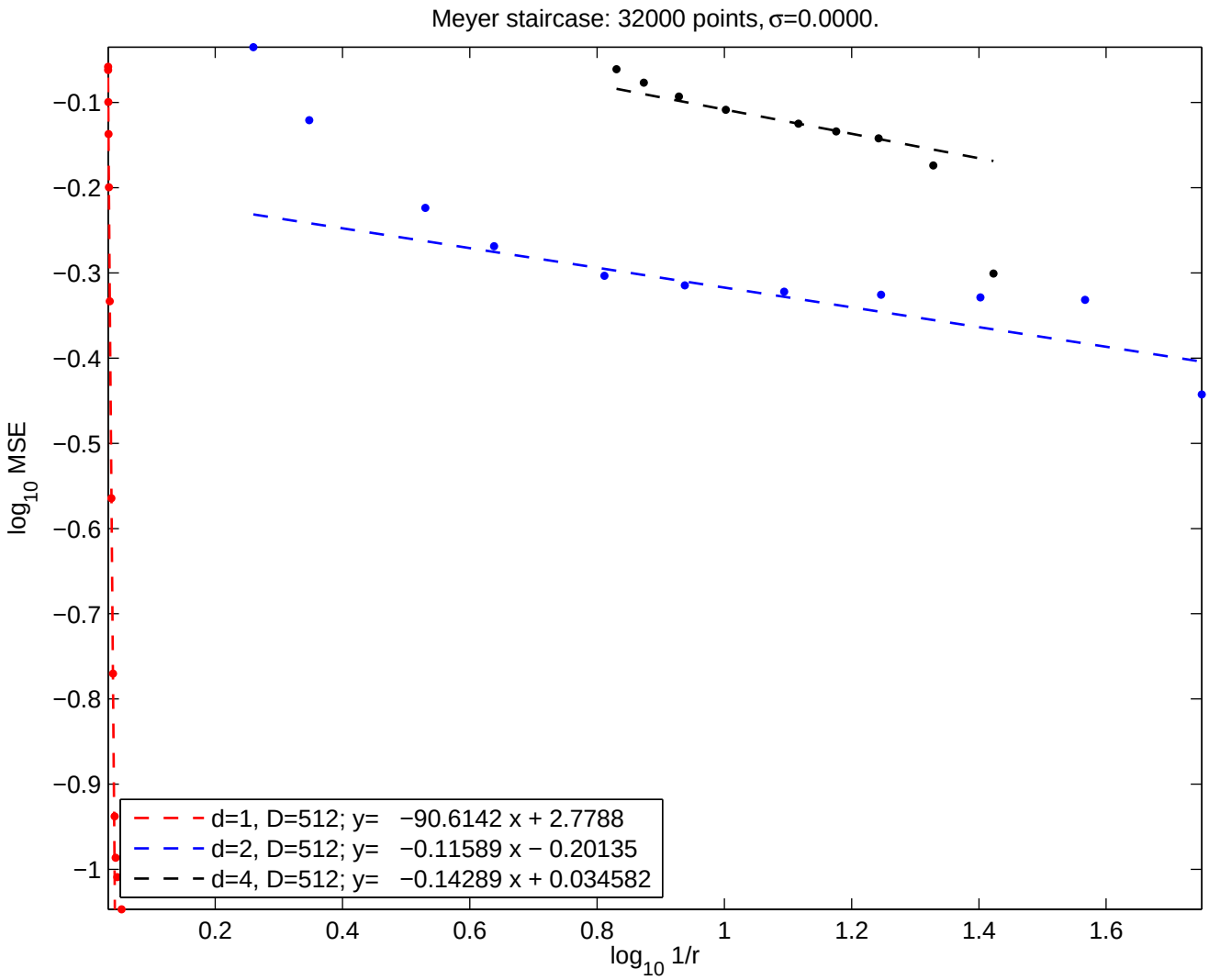
P. N. Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces. In *Proceedings of the 4th annual ACM-SIAM Symposium on Discrete algorithms*, pages 311–321. Society for Industrial and Applied Mathematics, 1993.

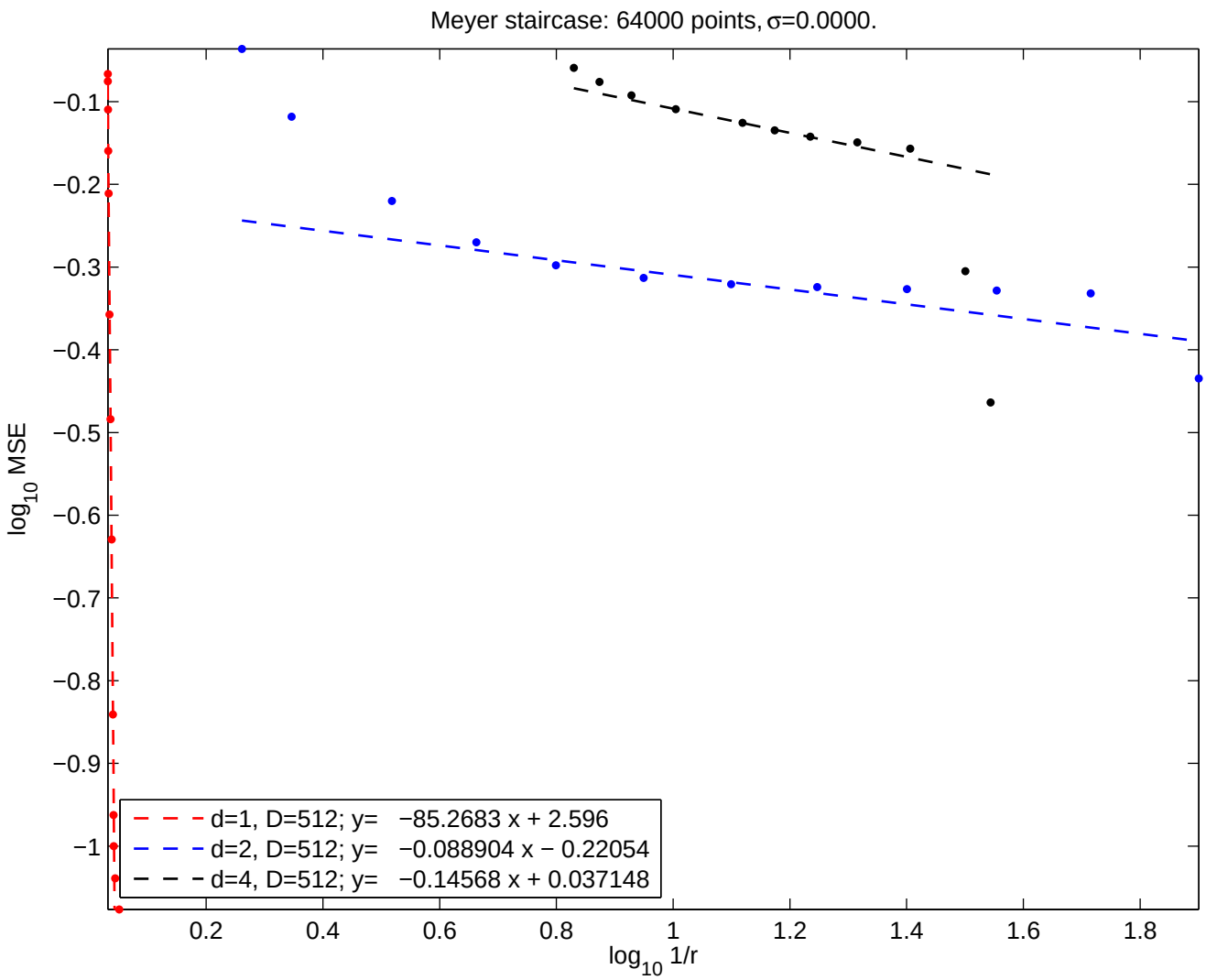
T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Randomized hybrid linear modeling by local best-fit flats. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1927–1934. IEEE, 2010.

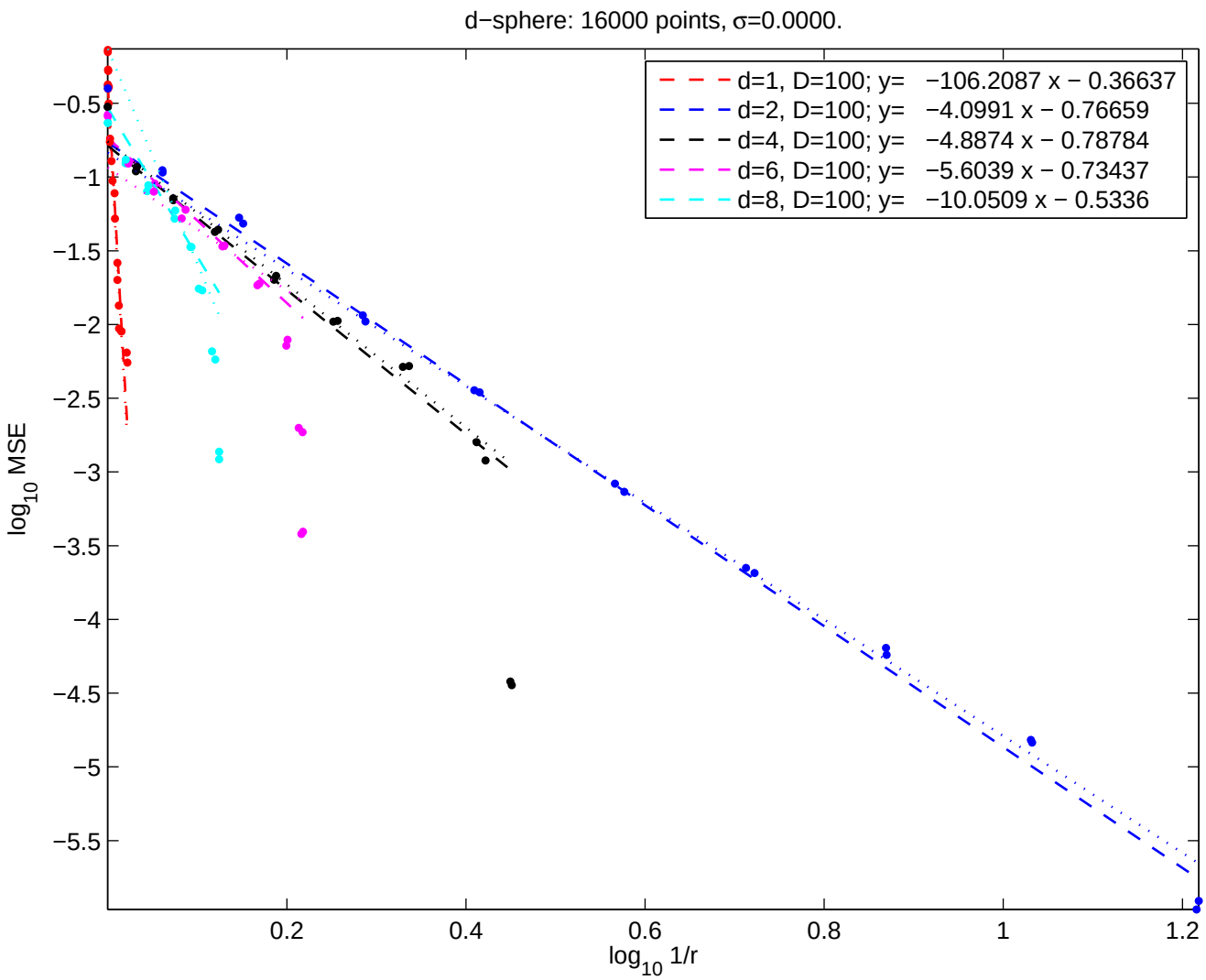
Z. Zhang and H. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *SIAM Journal of Scientific Computing*, 26:313–338, 2002.

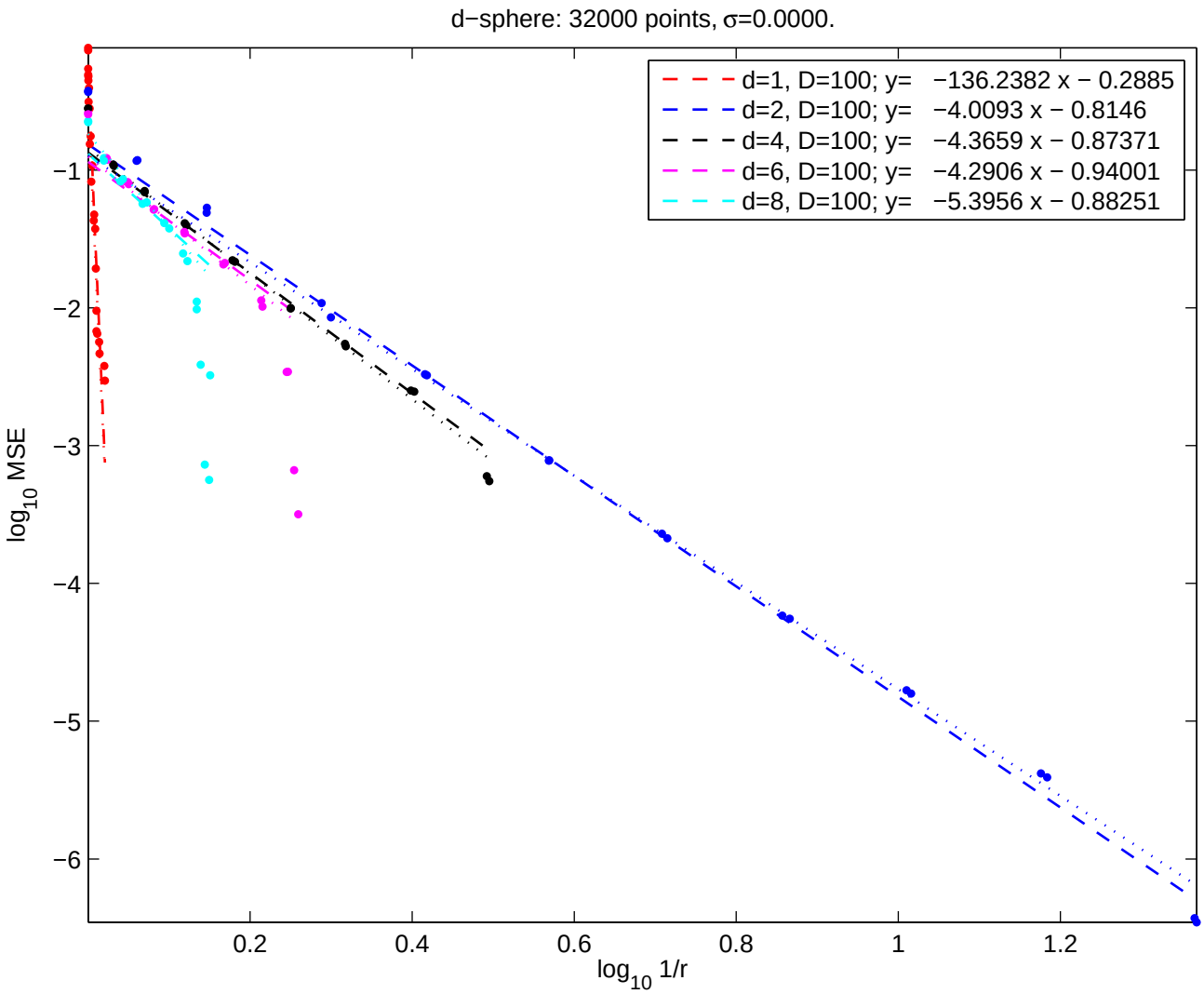
L. Zwald and G. Blanchard. On the convergence of eigenspaces in kernel principal component analysis. In *Advances in NIPS 18*, pages 1649–1656. MIT Press, Cambridge, MA, 2006.

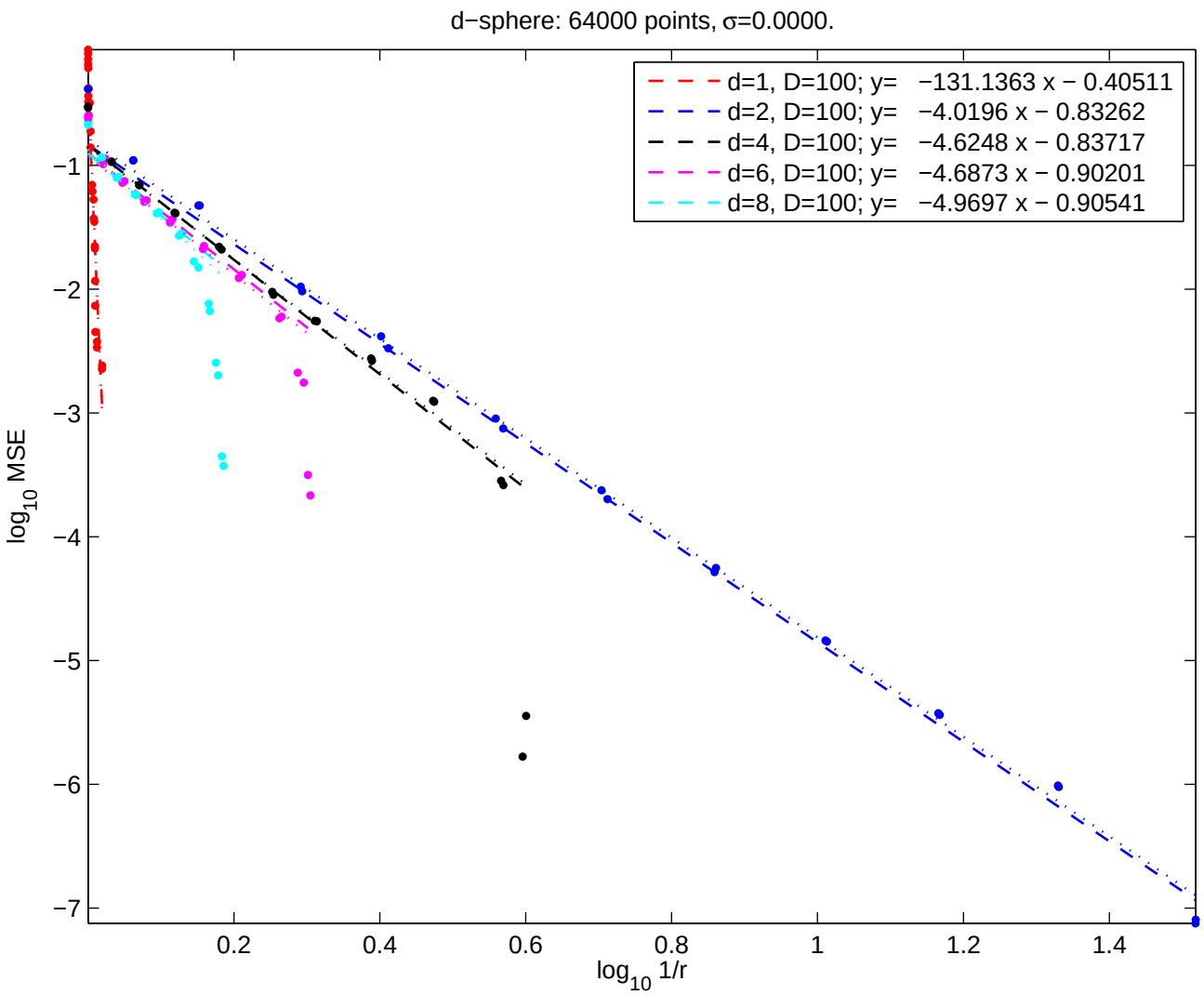

Meyer staircase: 16000 points, σ=0.0000.
log10 MSE
log10 1/r
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
−0.8
−0.9
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
d=1, D=512; y= −88.2978 x + 2.6938
d=2, D=512; y= −0.15768 x − 0.16801
d=4, D=512; y= −0.14227 x + 0.034857

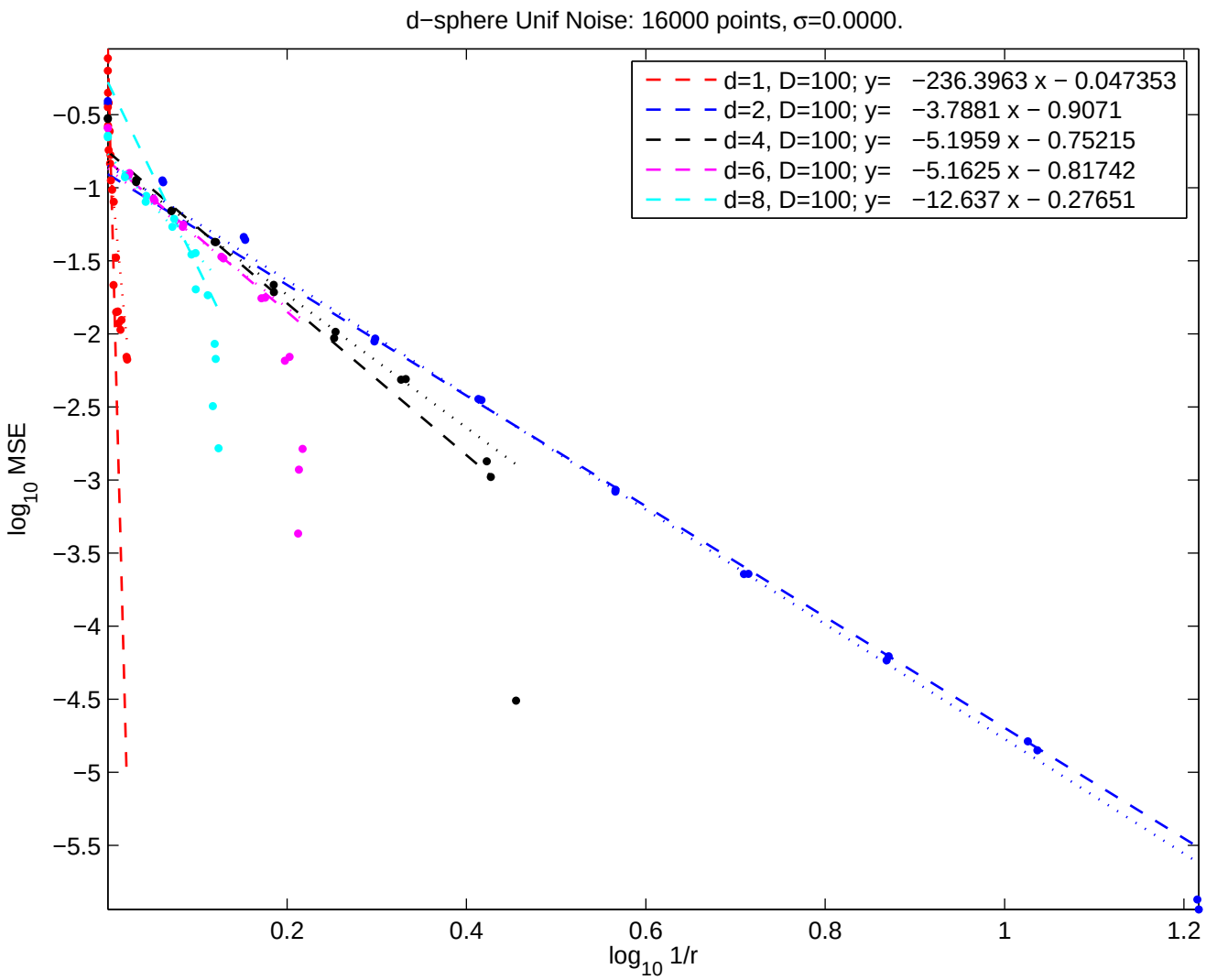

Meyer staircase: 32000 points, σ=0.0000.
log10 MSE
log10 1/r
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
−0.8
−0.9
−1
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
d=1, D=512; y= −90.6142 x + 2.7788
d=2, D=512; y= −0.11589 x − 0.20135
d=4, D=512; y= −0.14289 x + 0.034582

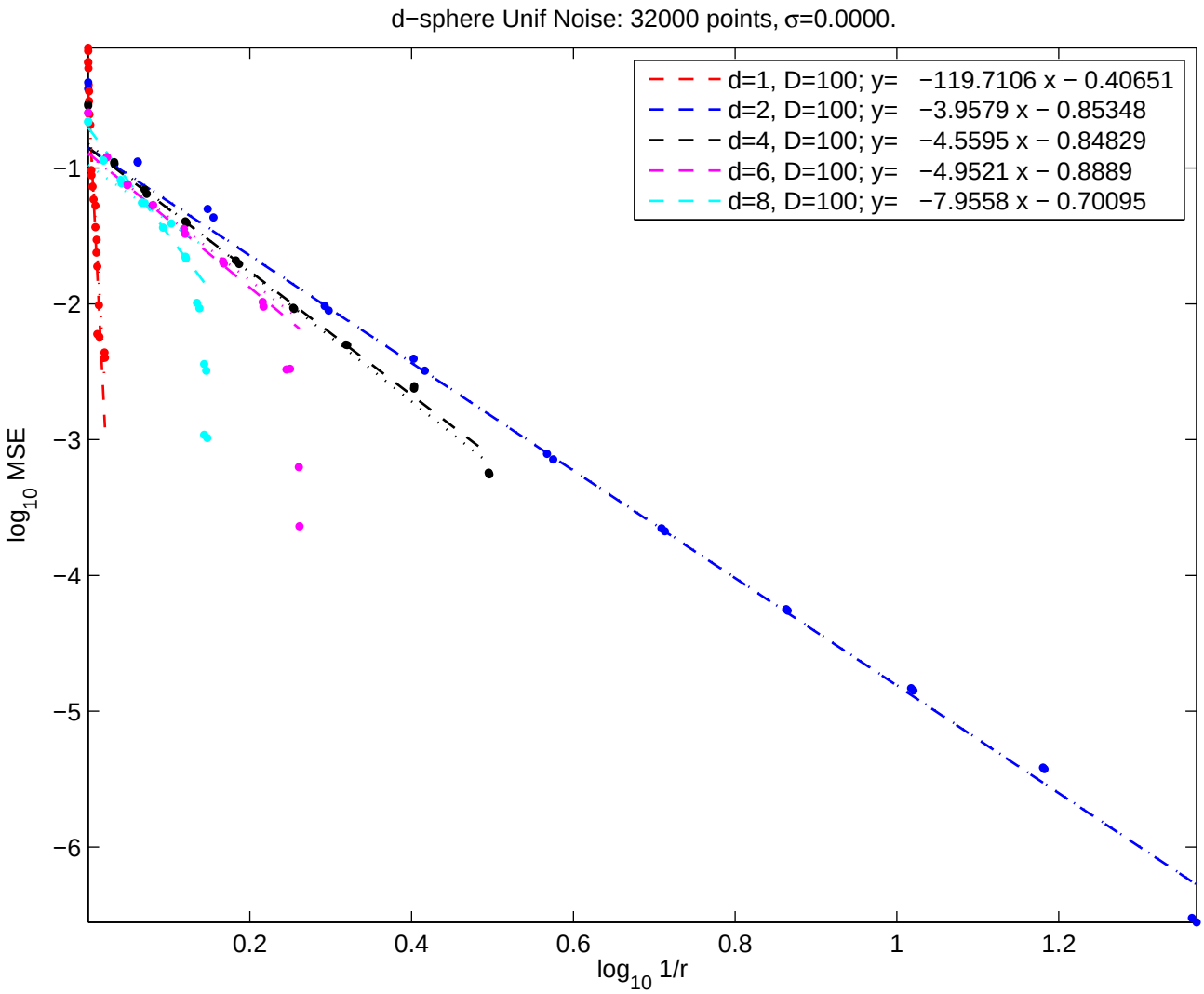

Meyer staircase: 64000 points, σ=0.0000.
log10 MSE
log10 1/r
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
−0.8
−0.9
−1
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
d=1, D=512; y= −85.2683 x + 2.596
d=2, D=512; y= −0.088904 x − 0.22054
d=4, D=512; y= −0.14568 x + 0.037148

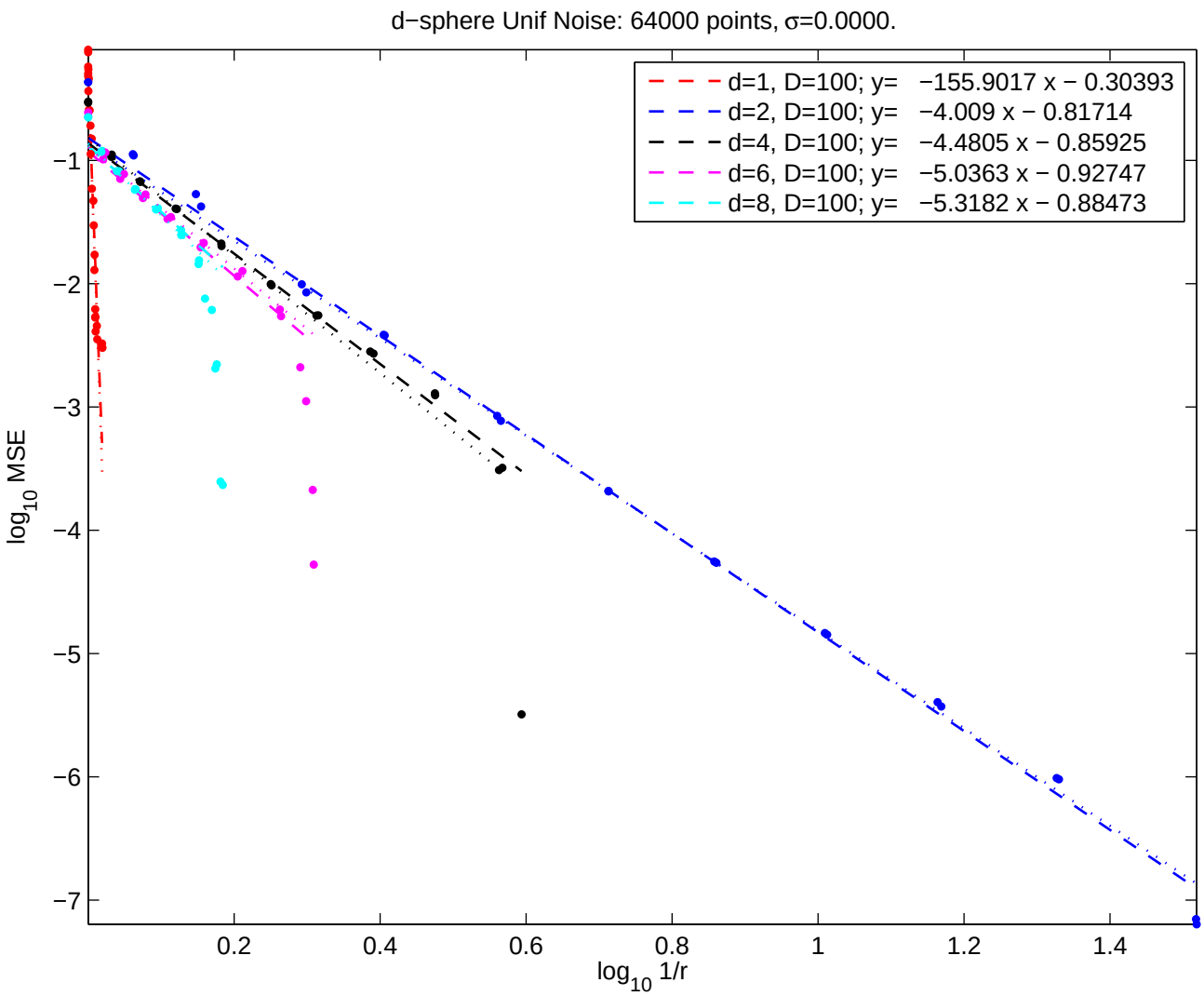

d−sphere: 16000 points, σ=0.0000.
d=1, D=100; y= −106.2087 x − 0.36637
d=2, D=100; y= −4.0991 x − 0.76659
d=4, D=100; y= −4.8874 x − 0.78784
d=6, D=100; y= −5.6039 x − 0.73437
d=8, D=100; y= −10.0509 x − 0.5336
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
−4
−4.5
−5
−5.5
0.2
0.4
0.6
0.8
1
1.2

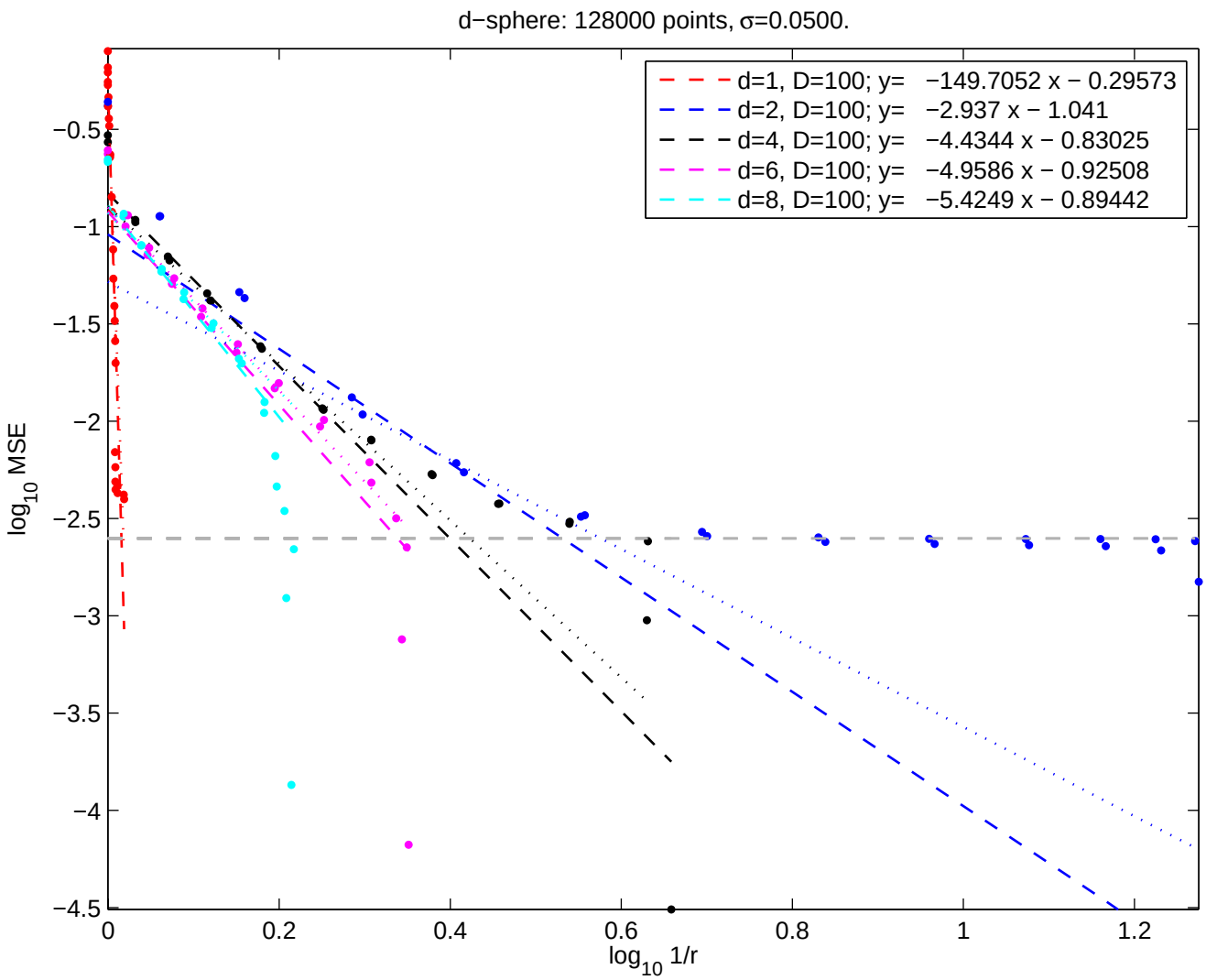

d−sphere: 32000 points, σ=0.0000.
d=1, D=100; y= −136.2382 x − 0.2885
d=2, D=100; y= −4.0093 x − 0.8146
d=4, D=100; y= −4.3659 x − 0.87371
d=6, D=100; y= −4.2906 x − 0.94001
d=8, D=100; y= −5.3956 x − 0.88251
log10 MSE
log10 1/r
−1
−2
−3
−4
−5
−6
0.2
0.4
0.6
0.8
1
1.2

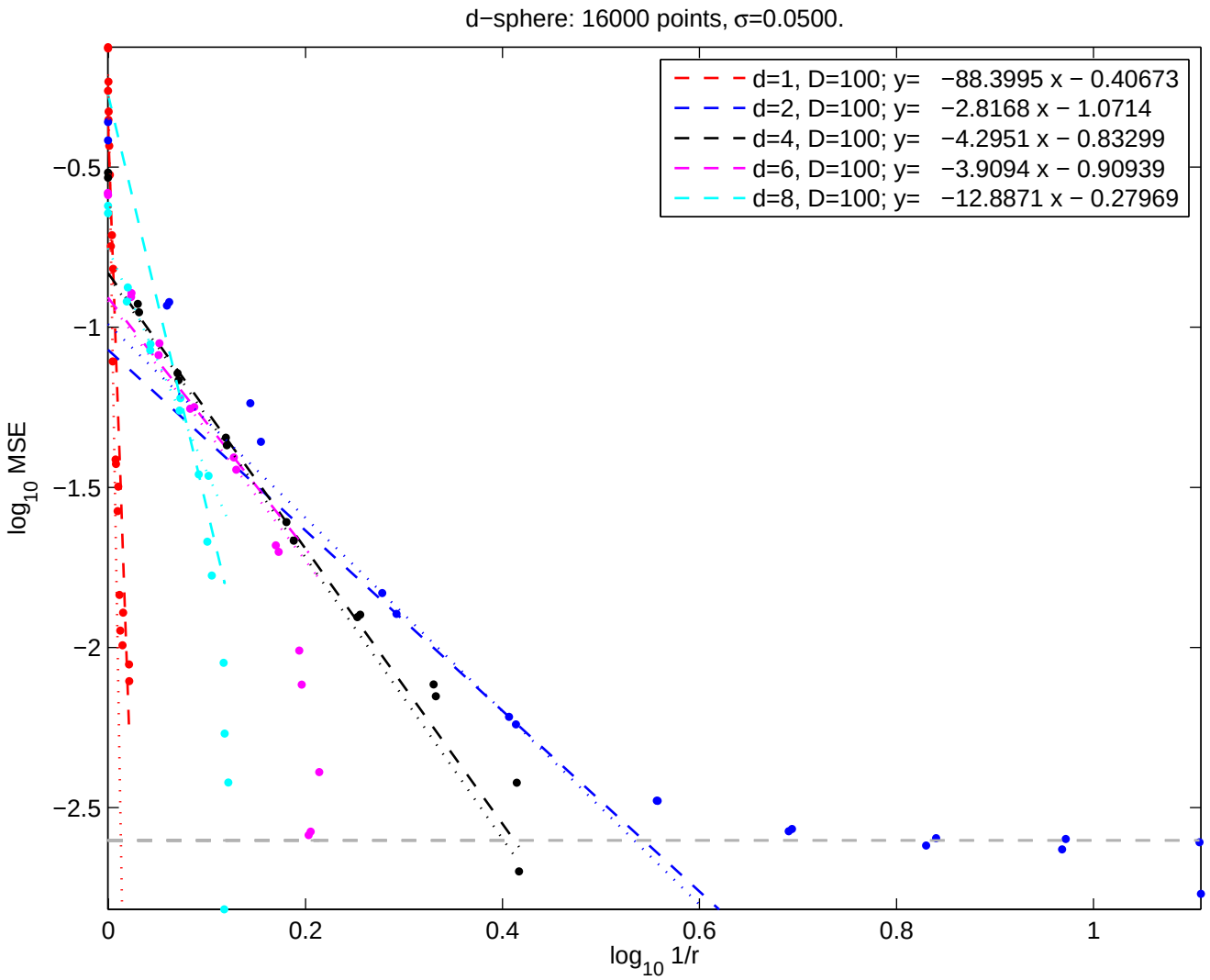

d−sphere: 64000 points, σ=0.0000.
d=1, D=100; y= −131.1363 x − 0.40511
d=2, D=100; y= −4.0196 x − 0.83262
d=4, D=100; y= −4.6248 x − 0.83717
d=6, D=100; y= −4.6873 x − 0.90201
d=8, D=100; y= −4.9697 x − 0.90541
log10 MSE
log10 1/r
−1
−2
−3
−4
−5
−6
−7
0.2
0.4
0.6
0.8
1
1.2
1.4

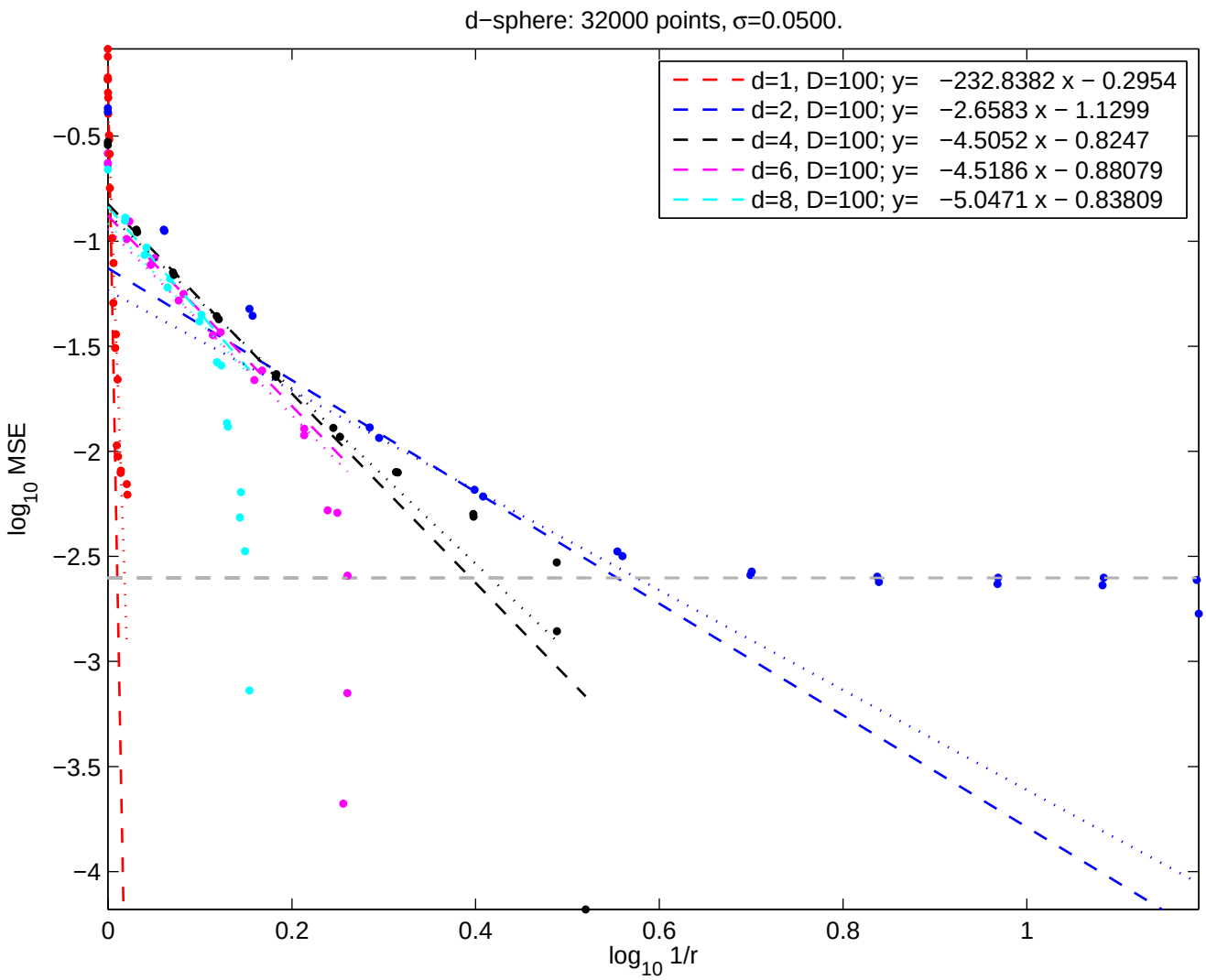

d−sphere Unif Noise: 16000 points, σ=0.0000.
d=1, D=100; y= −236.3963 x − 0.047353
d=2, D=100; y= −3.7881 x − 0.9071
d=4, D=100; y= −5.1959 x − 0.75215
d=6, D=100; y= −5.1625 x − 0.81742
d=8, D=100; y= −12.637 x − 0.27651
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
−4
−4.5
−5
−5.5
0.2
0.4
0.6
0.8
1
1.2

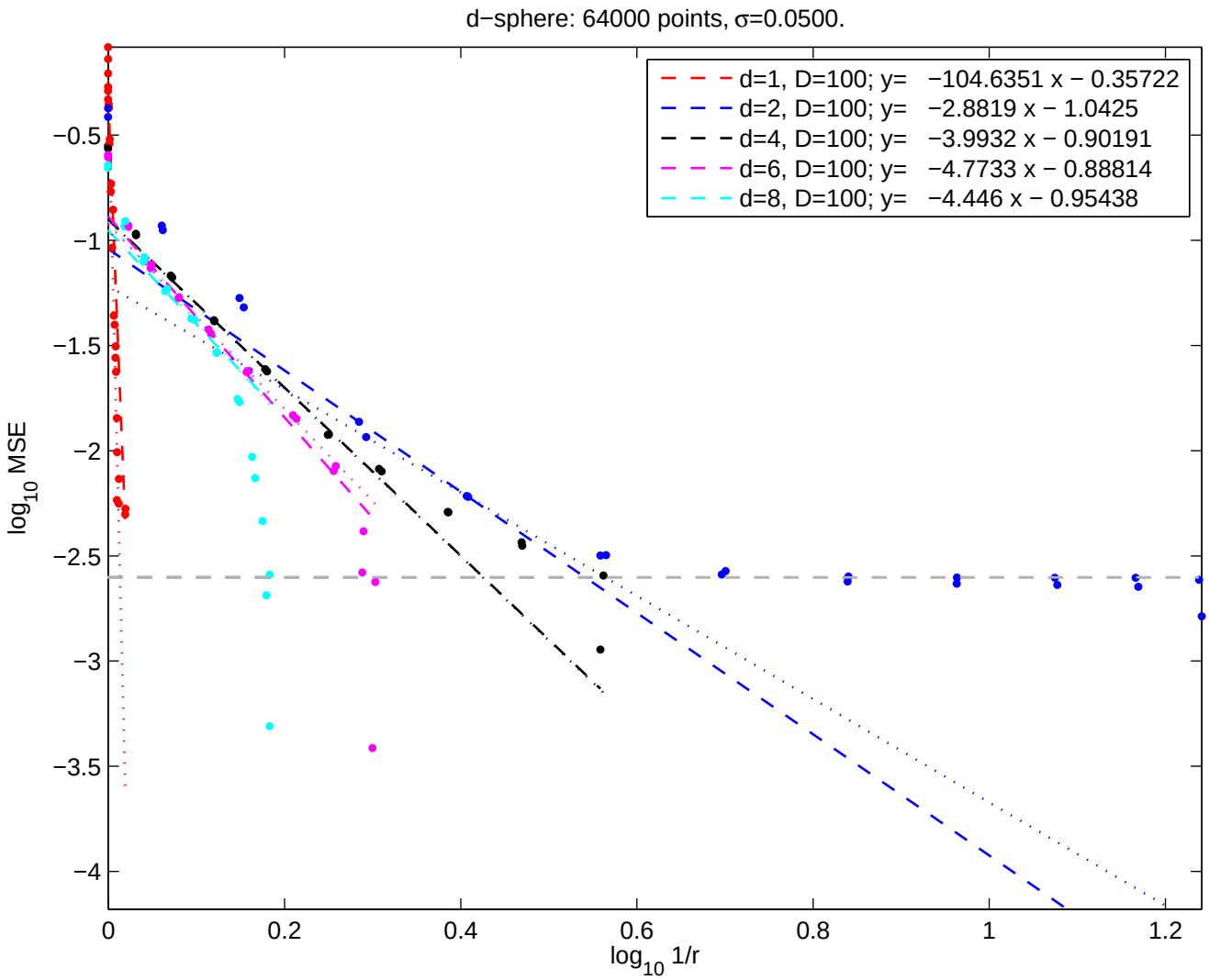

d−sphere Unif Noise: 32000 points, σ=0.0000.
d=1, D=100; y= −119.7106 x − 0.40651
d=2, D=100; y= −3.9579 x − 0.85348
d=4, D=100; y= −4.5595 x − 0.84829
d=6, D=100; y= −4.9521 x − 0.8889
d=8, D=100; y= −7.9558 x − 0.70095
log10 MSE
log10 1/r
−1
−2
−3
−4
−5
−6
0.2
0.4
0.6
0.8
1
1.2

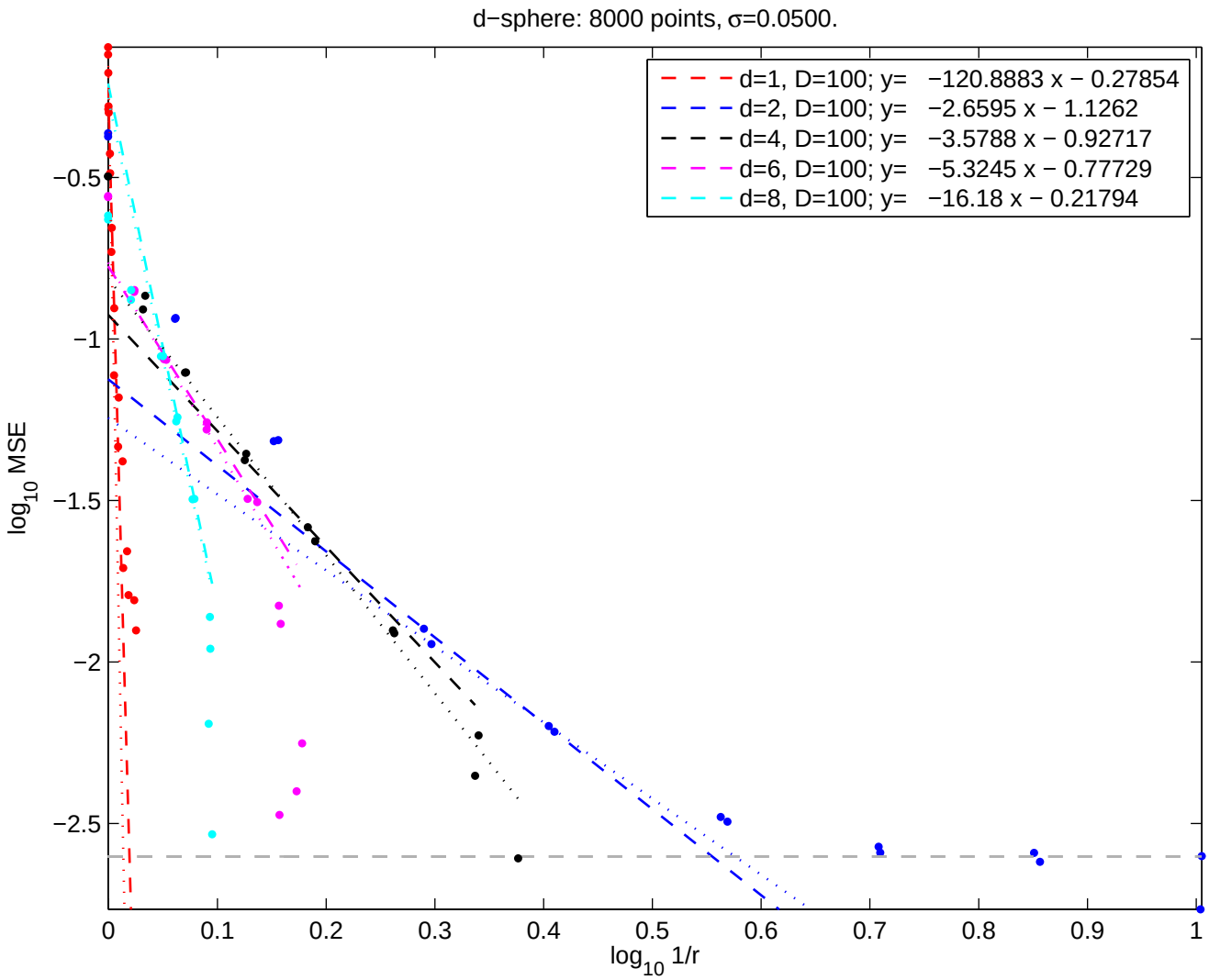

d−sphere Unif Noise: 64000 points, σ=0.0000.
d=1, D=100; y= −155.9017 x − 0.30393
d=2, D=100; y= −4.009 x − 0.81714
d=4, D=100; y= −4.4805 x − 0.85925
d=6, D=100; y= −5.0363 x − 0.92747
d=8, D=100; y= −5.3182 x − 0.88473
log10 MSE
log10 1/r
−1
−2
−3
−4
−5
−6
−7
0.2
0.4
0.6
0.8
1
1.2
1.4

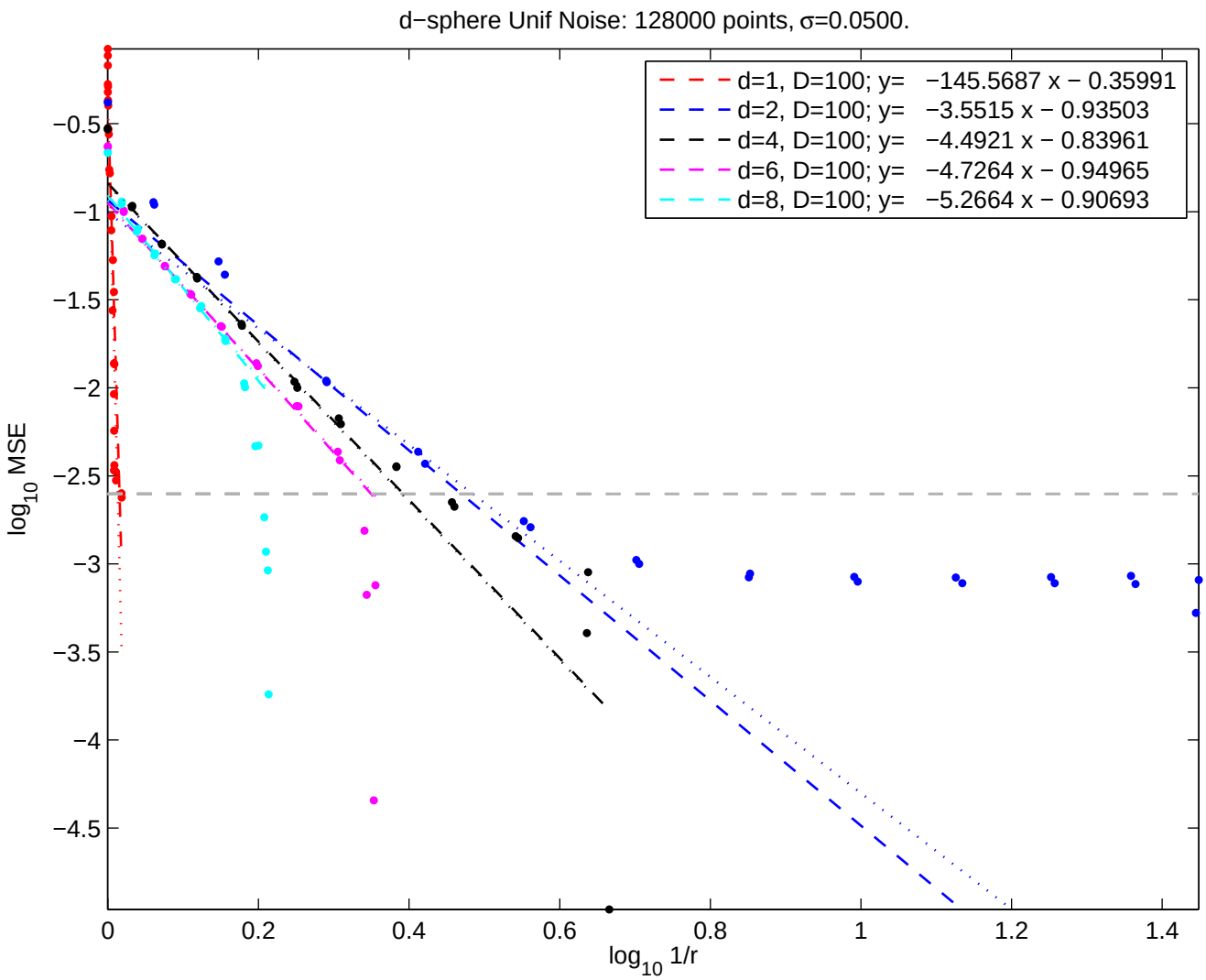

d−sphere: 128000 points, σ=0.0500.
d=1, D=100; y= −149.7052 x − 0.29573
d=2, D=100; y= −2.937 x − 1.041
d=4, D=100; y= −4.4344 x − 0.83025
d=6, D=100; y= −4.9586 x − 0.92508
d=8, D=100; y= −5.4249 x − 0.89442
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
−4
−4.5
0
0.2
0.4
0.6
0.8
1
1.2

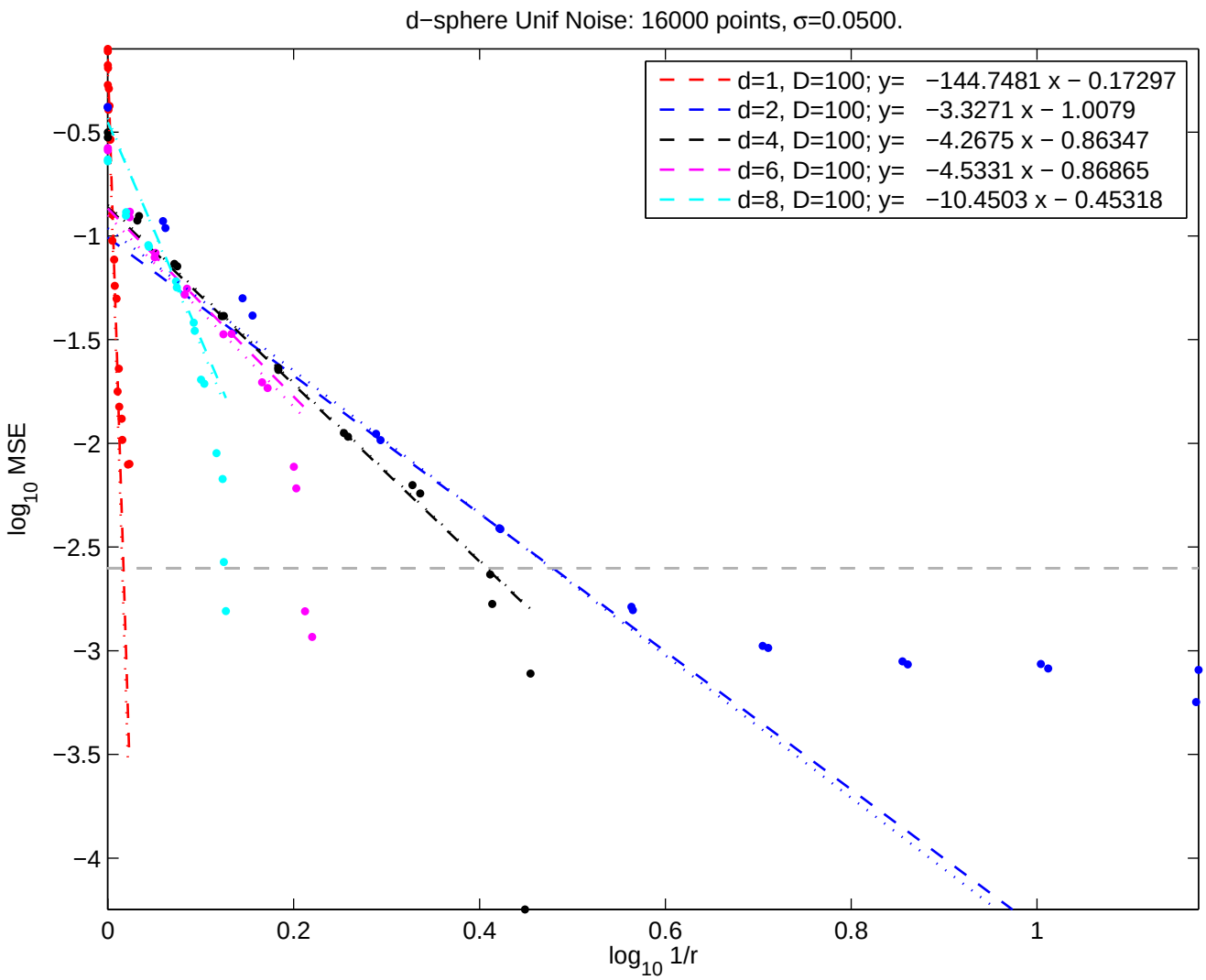

d−sphere: 16000 points, σ=0.0500.
d=1, D=100; y= −88.3995 x − 0.40673
d=2, D=100; y= −2.8168 x − 1.0714
d=4, D=100; y= −4.2951 x − 0.83299
d=6, D=100; y= −3.9094 x − 0.90939
d=8, D=100; y= −12.8871 x − 0.27969
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
0
0.2
0.4
0.6
0.8
1

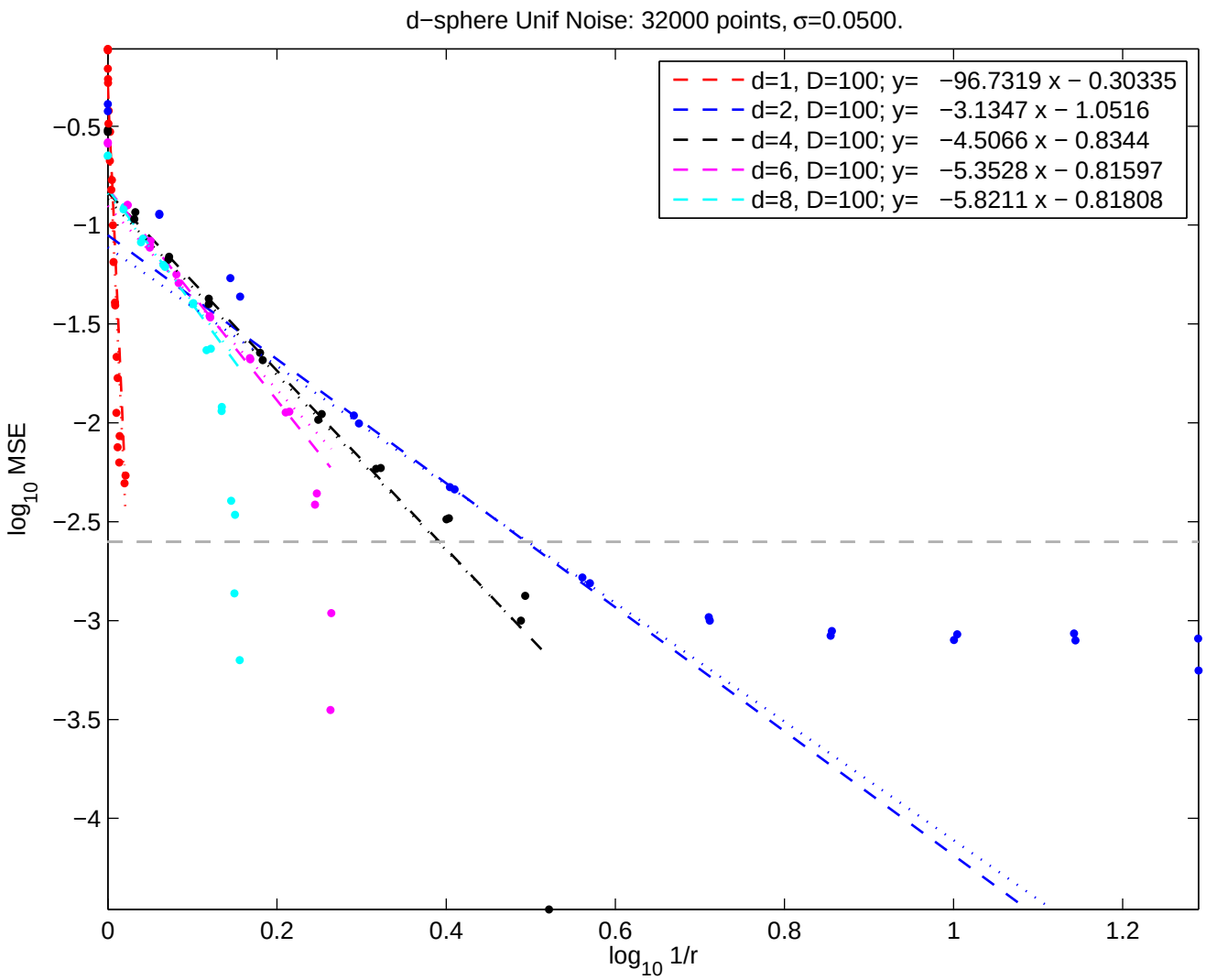

d−sphere: 32000 points, σ=0.0500.
d=1, D=100; y= −232.8382 x − 0.2954
d=2, D=100; y= −2.6583 x − 1.1299
d=4, D=100; y= −4.5052 x − 0.8247
d=6, D=100; y= −4.5186 x − 0.88079
d=8, D=100; y= −5.0471 x − 0.83809
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
−4
0
0.2
0.4
0.6
0.8
1

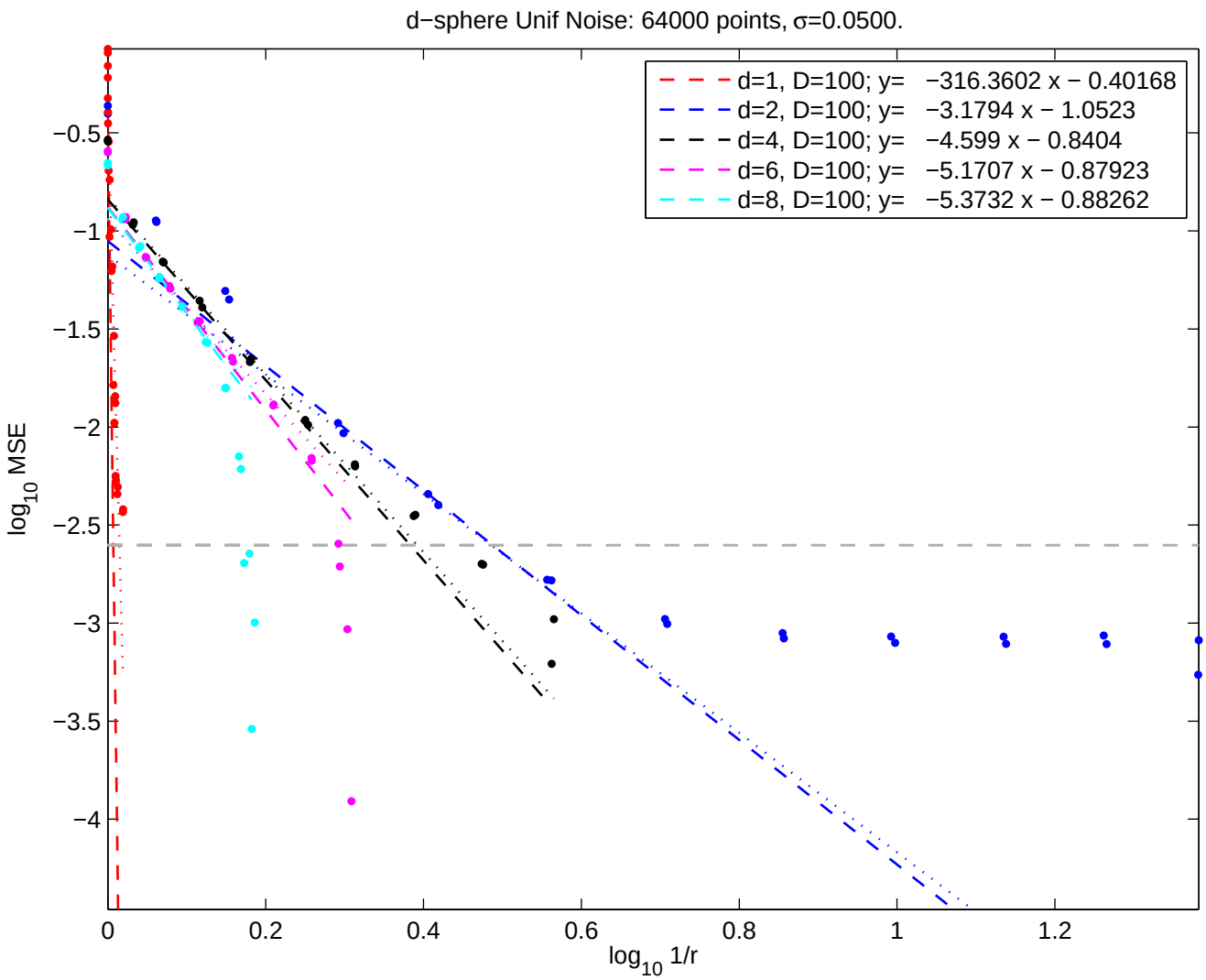

d−sphere: 64000 points, σ=0.0500.
d=1, D=100; y= −104.6351 x − 0.35722
d=2, D=100; y= −2.8819 x − 1.0425
d=4, D=100; y= −3.9932 x − 0.90191
d=6, D=100; y= −4.7733 x − 0.88814
d=8, D=100; y= −4.446 x − 0.95438
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
−4
0
0.2
0.4
0.6
0.8
1
1.2

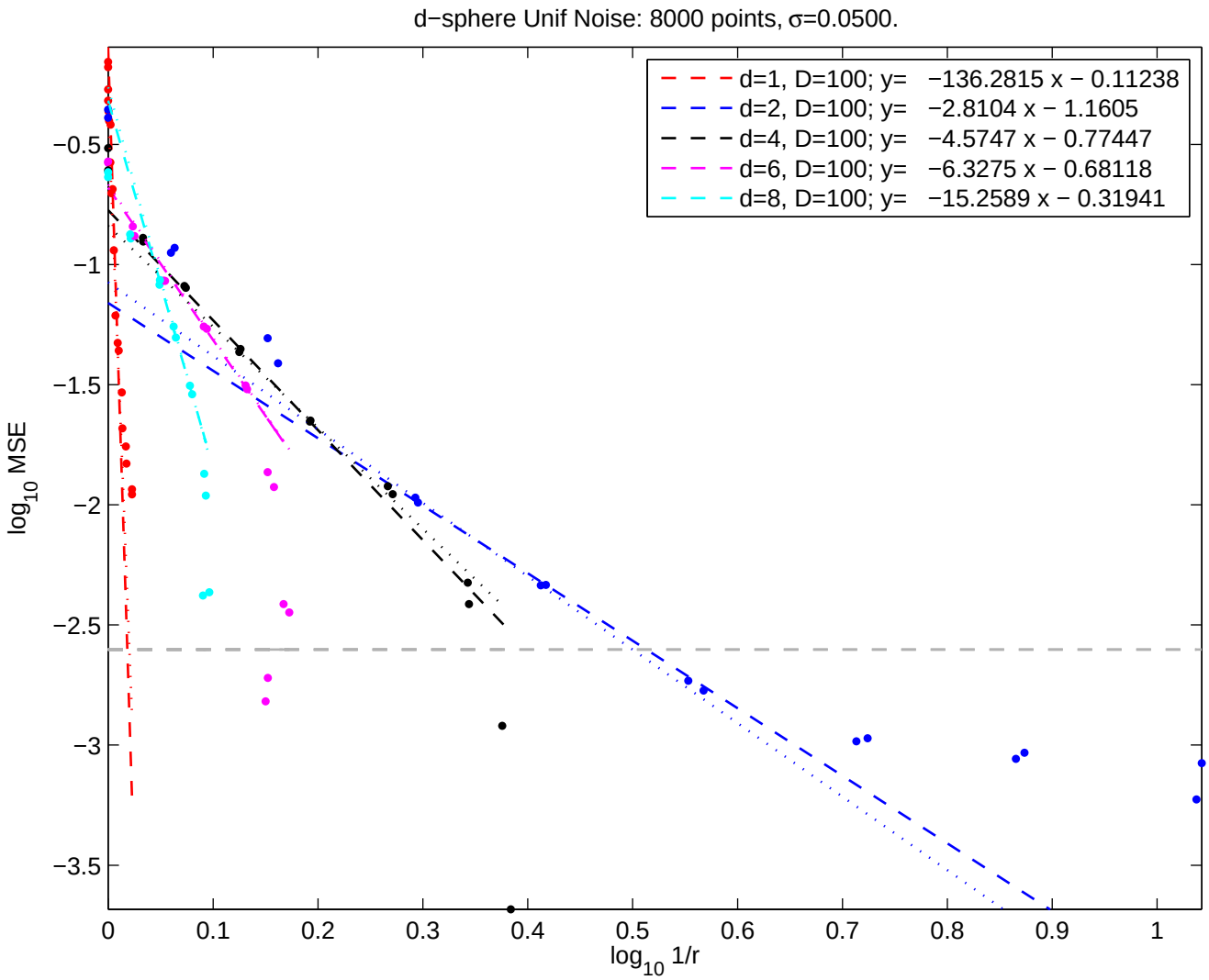

d−sphere: 8000 points, σ=0.0500.
d=1, D=100; y= −120.8883 x − 0.27854
d=2, D=100; y= −2.6595 x − 1.1262
d=4, D=100; y= −3.5788 x − 0.92717
d=6, D=100; y= −5.3245 x − 0.77729
d=8, D=100; y= −16.18 x − 0.21794
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

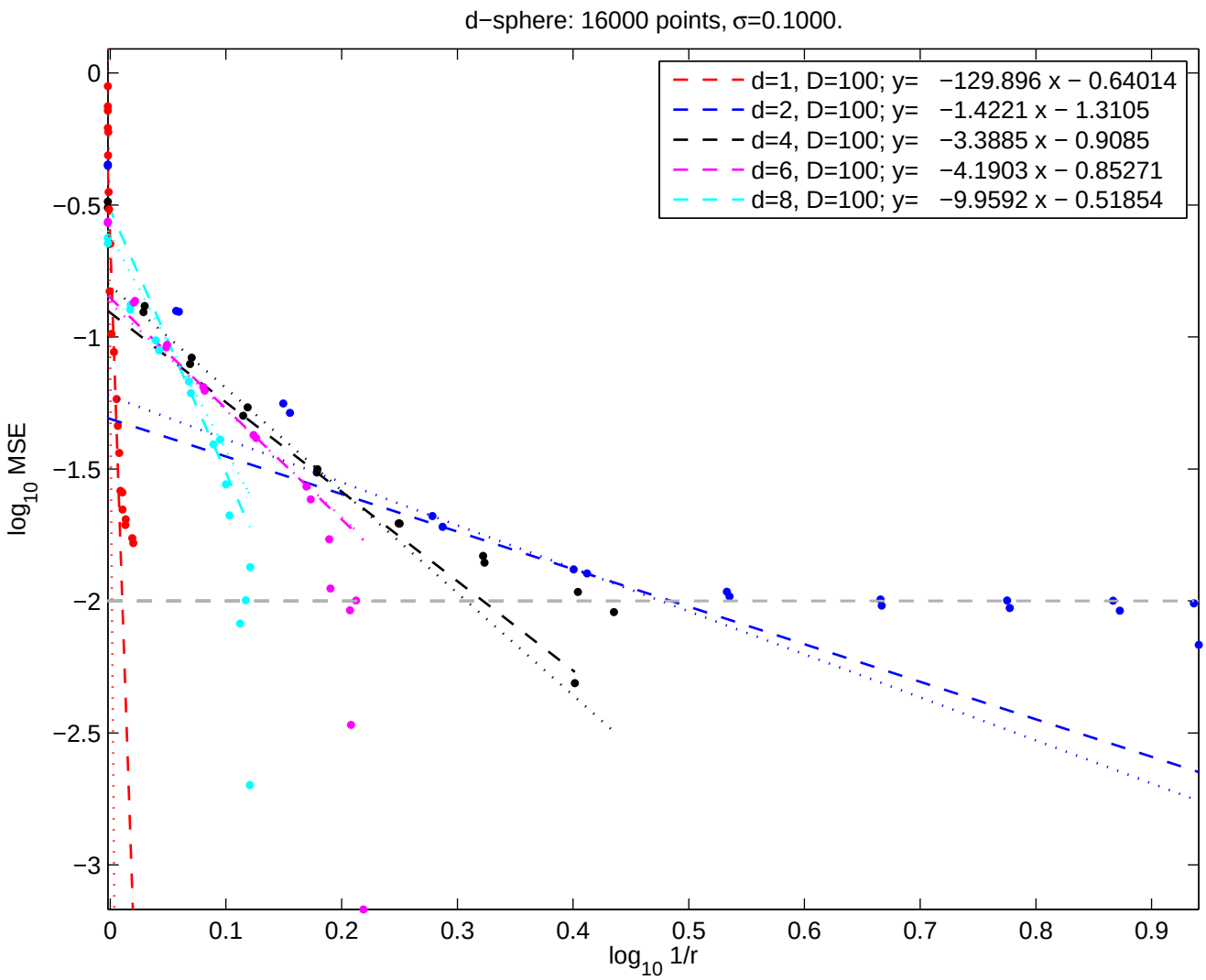

d−sphere Unif Noise: 128000 points, σ=0.0500.
d=1, D=100; y= −145.5687 x − 0.35991
d=2, D=100; y= −3.5515 x − 0.93503
d=4, D=100; y= −4.4921 x − 0.83961
d=6, D=100; y= −4.7264 x − 0.94965
d=8, D=100; y= −5.2664 x − 0.90693
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
−4
−4.5
0
0.2
0.4
0.6
0.8
1
1.2
1.4

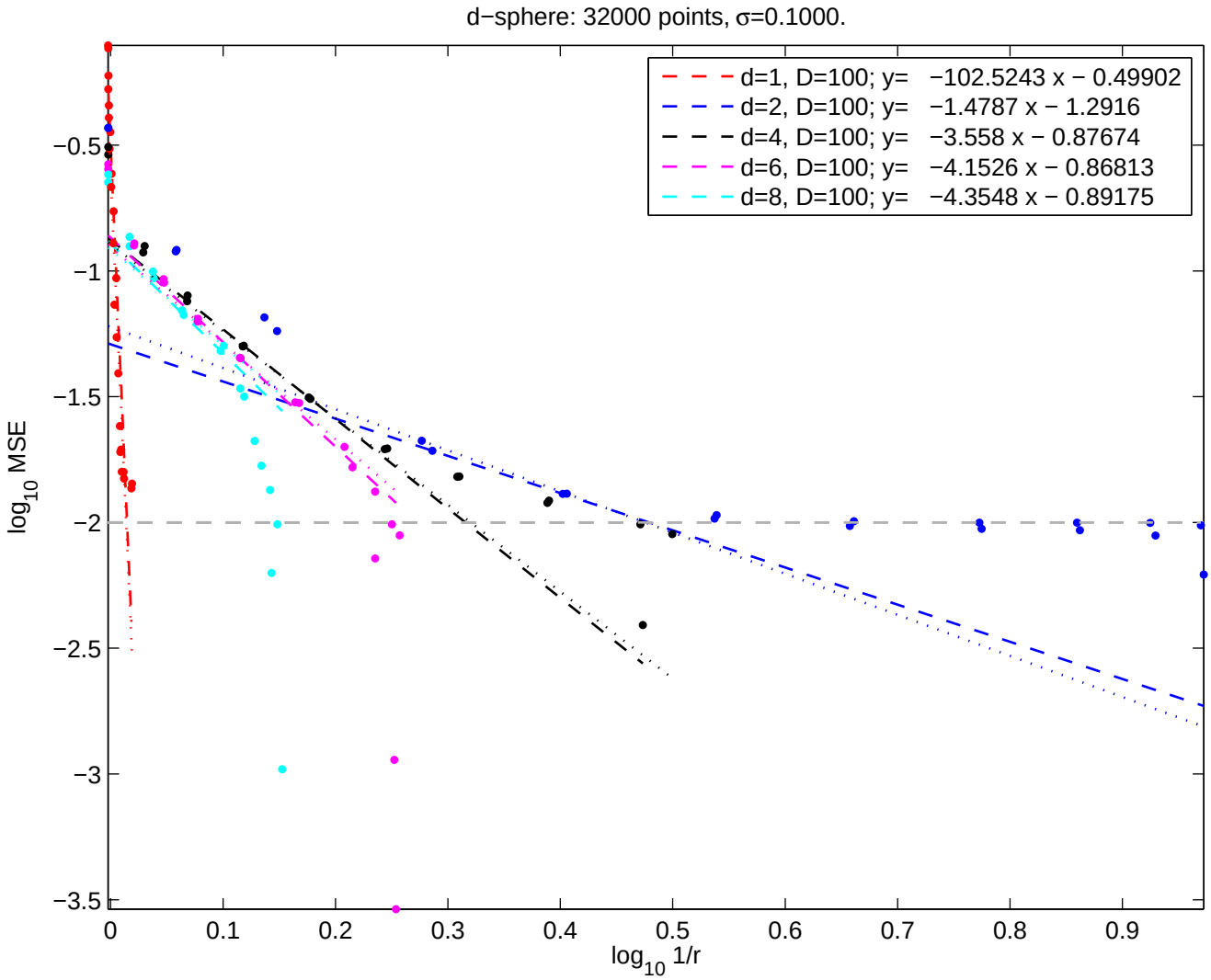

d−sphere Unif Noise: 16000 points, σ=0.0500.
d=1, D=100; y= −144.7481 x − 0.17297
d=2, D=100; y= −3.3271 x − 1.0079
d=4, D=100; y= −4.2675 x − 0.86347
d=6, D=100; y= −4.5331 x − 0.86865
d=8, D=100; y= −10.4503 x − 0.45318
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
−4
0
0.2
0.4
0.6
0.8
1

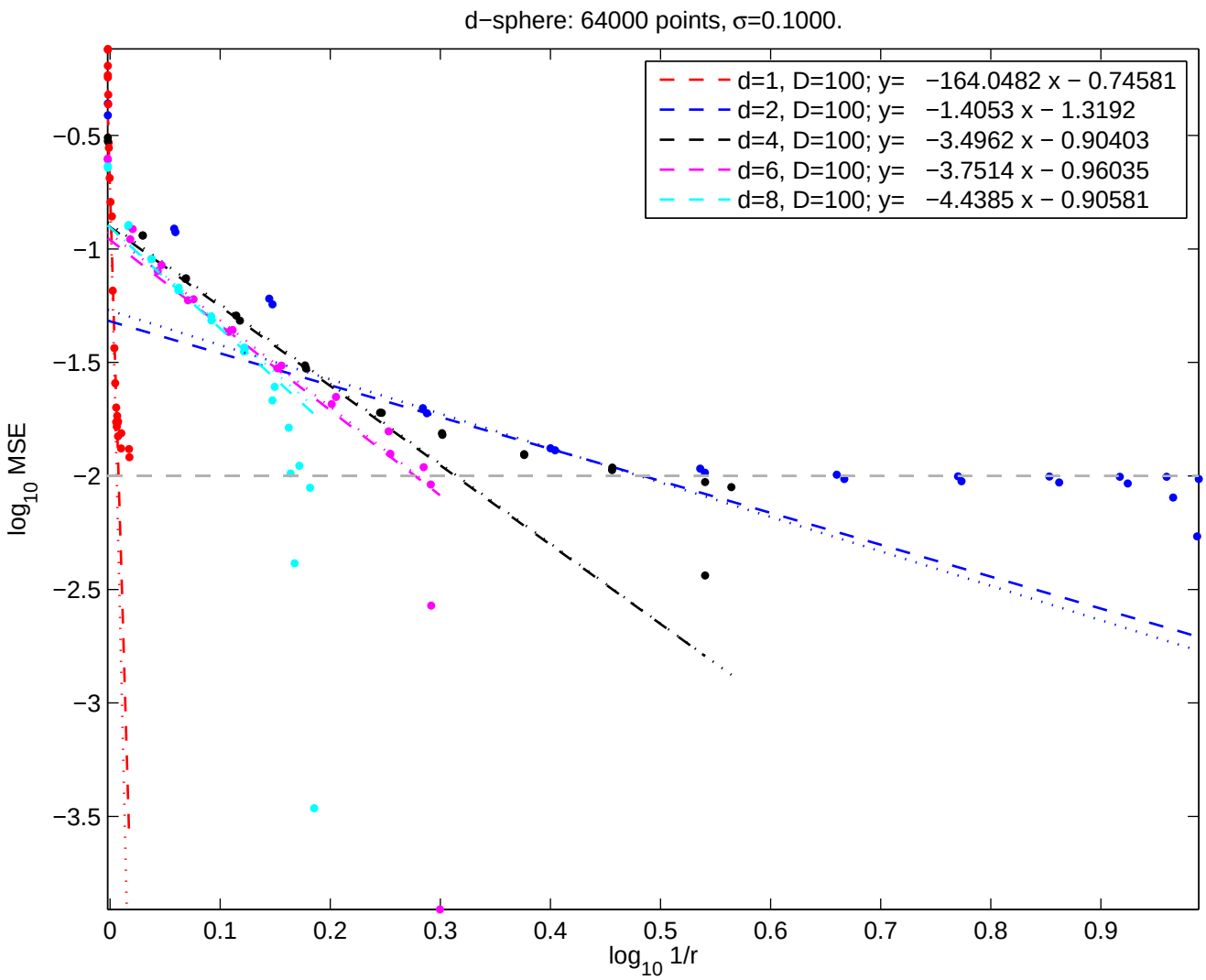

d−sphere Unif Noise: 32000 points, σ=0.0500.
d=1, D=100; y= −96.7319 x − 0.30335
d=2, D=100; y= −3.1347 x − 1.0516
d=4, D=100; y= −4.5066 x − 0.8344
d=6, D=100; y= −5.3528 x − 0.81597
d=8, D=100; y= −5.8211 x − 0.81808
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
−4
0
0.2
0.4
0.6
0.8
1
1.2

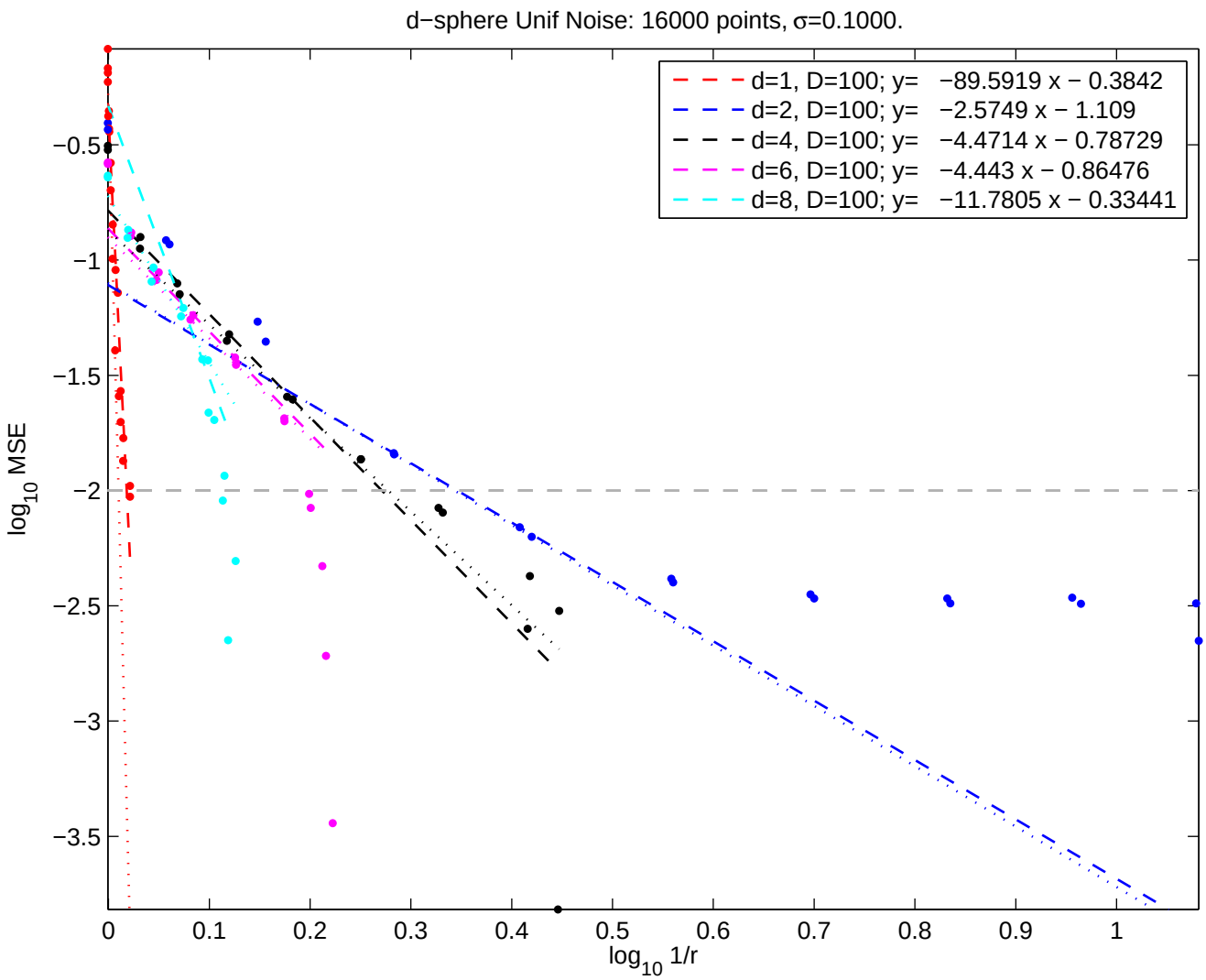

d−sphere Unif Noise: 64000 points, σ=0.0500.
d=1, D=100; y= −316.3602 x − 0.40168
d=2, D=100; y= −3.1794 x − 1.0523
d=4, D=100; y= −4.599 x − 0.8404
d=6, D=100; y= −5.1707 x − 0.87923
d=8, D=100; y= −5.3732 x − 0.88262
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
−4
0
0.2
0.4
0.6
0.8
1
1.2

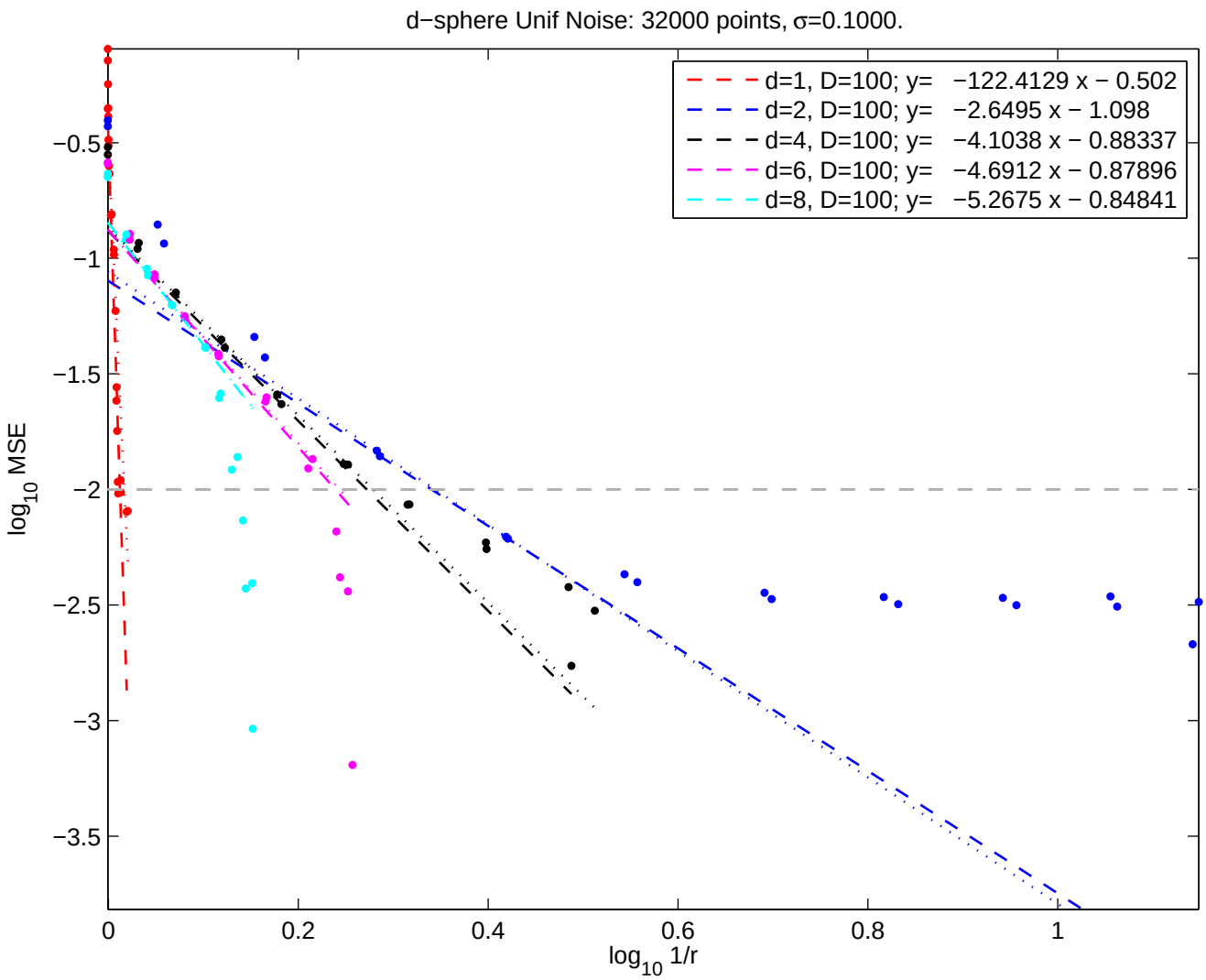

d−sphere Unif Noise: 8000 points, σ=0.0500.
d=1, D=100; y= −136.2815 x − 0.11238
d=2, D=100; y= −2.8104 x − 1.1605
d=4, D=100; y= −4.5747 x − 0.77447
d=6, D=100; y= −6.3275 x − 0.68118
d=8, D=100; y= −15.2589 x − 0.31941
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

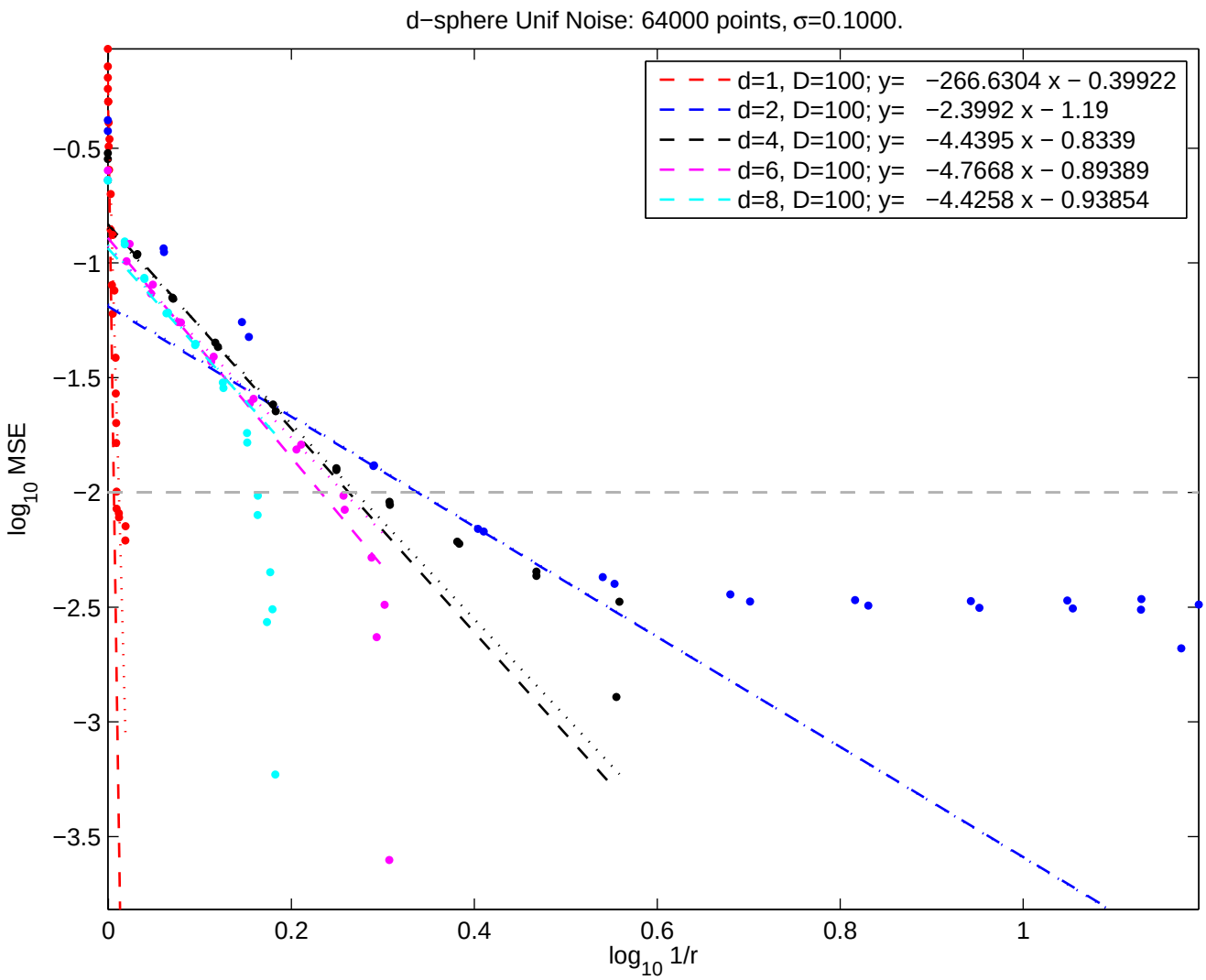

d−sphere: 16000 points, σ=0.1000.
d=1, D=100; y= −129.896 x − 0.64014
d=2, D=100; y= −1.4221 x − 1.3105
d=4, D=100; y= −3.3885 x − 0.9085
d=6, D=100; y= −4.1903 x − 0.85271
d=8, D=100; y= −9.9592 x − 0.51854
log10 MSE
log10 1/r
0
−0.5
−1
−1.5
−2
−2.5
−3
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

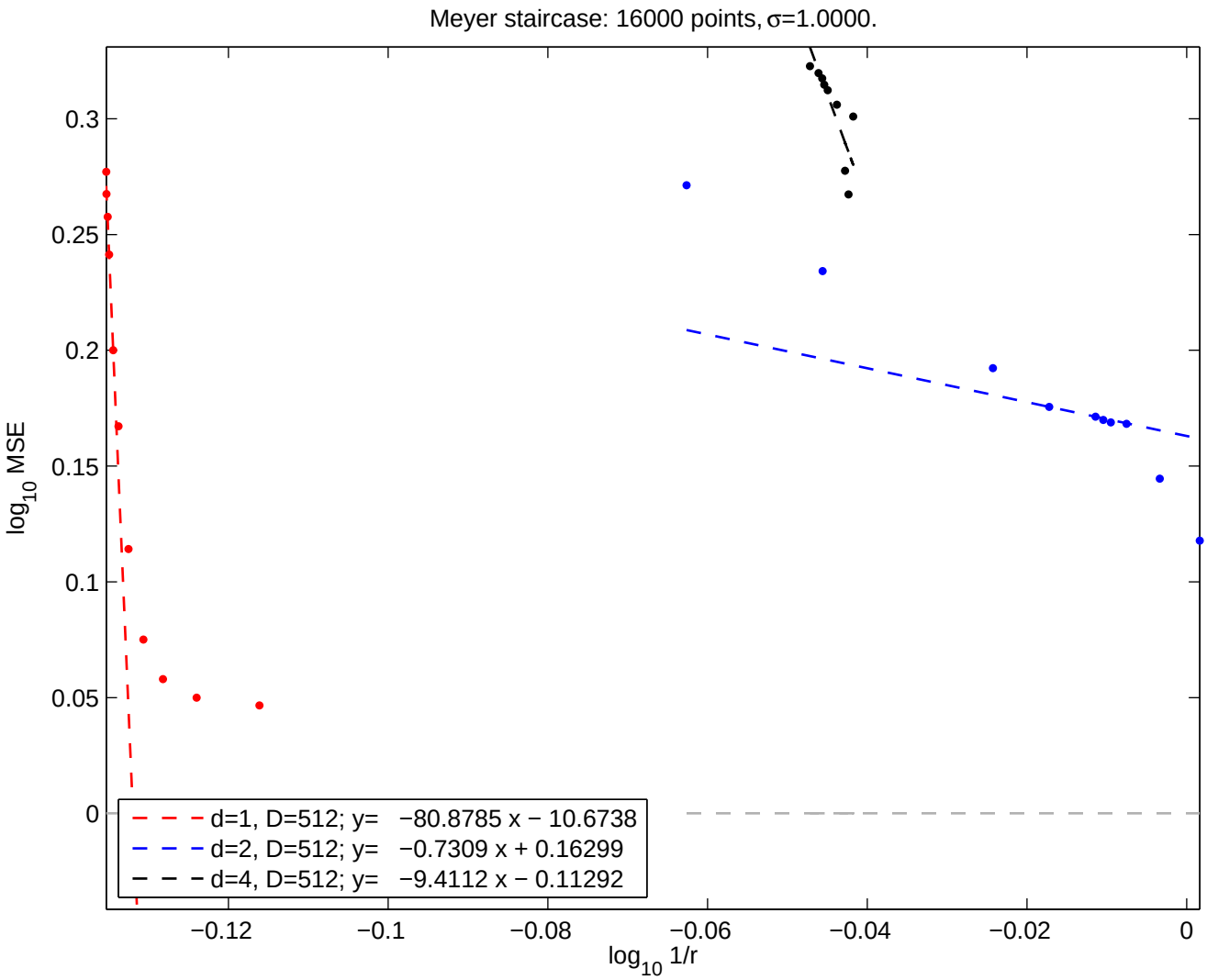

d−sphere: 32000 points, σ=0.1000.
d=1, D=100; y= −102.5243 x − 0.49902
d=2, D=100; y= −1.4787 x − 1.2916
d=4, D=100; y= −3.558 x − 0.87674
d=6, D=100; y= −4.1526 x − 0.86813
d=8, D=100; y= −4.3548 x − 0.89175
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

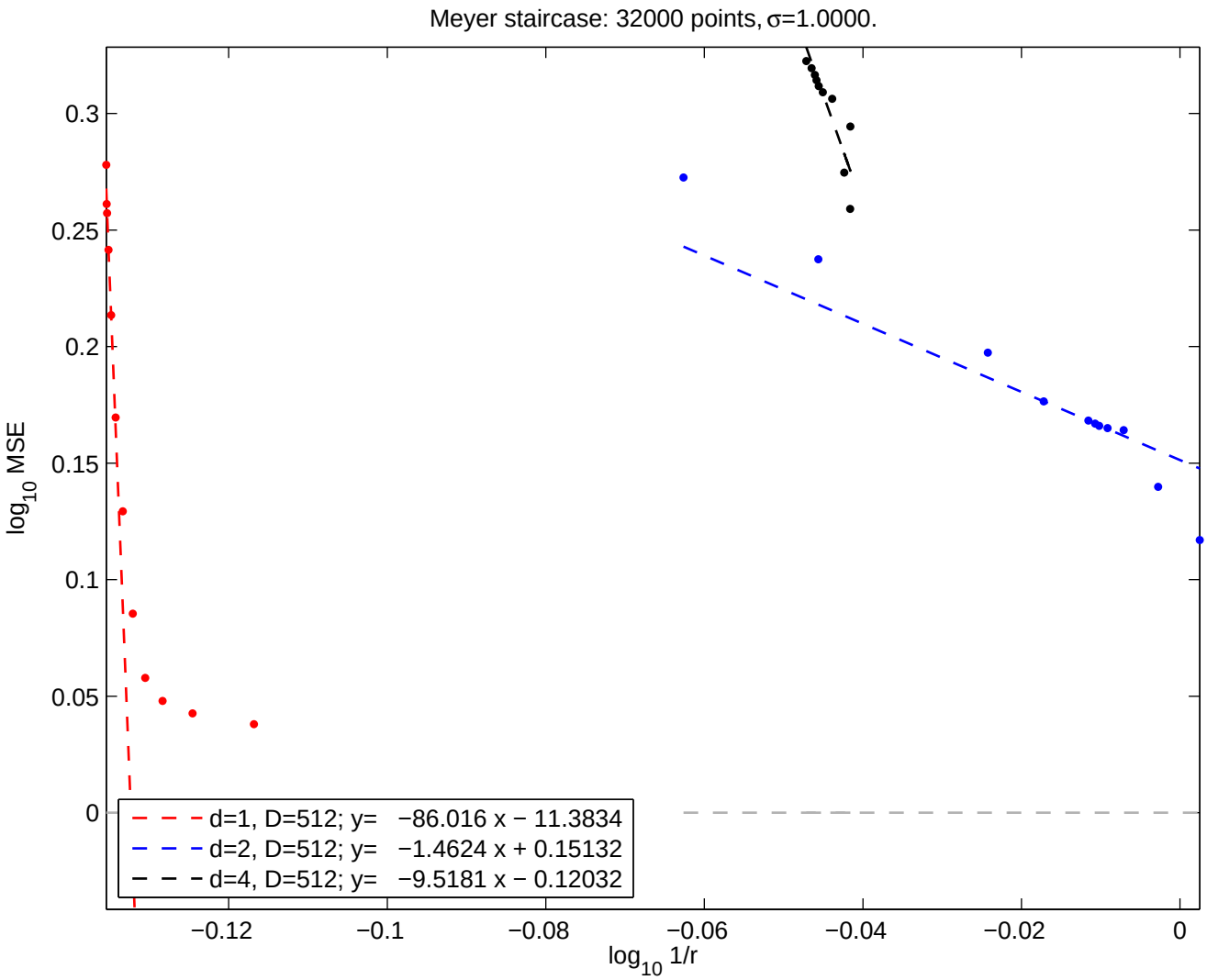

d−sphere: 64000 points, σ=0.1000.
d=1, D=100; y= −164.0482 x − 0.74581
d=2, D=100; y= −1.4053 x − 1.3192
d=4, D=100; y= −3.4962 x − 0.90403
d=6, D=100; y= −3.7514 x − 0.96035
d=8, D=100; y= −4.4385 x − 0.90581
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

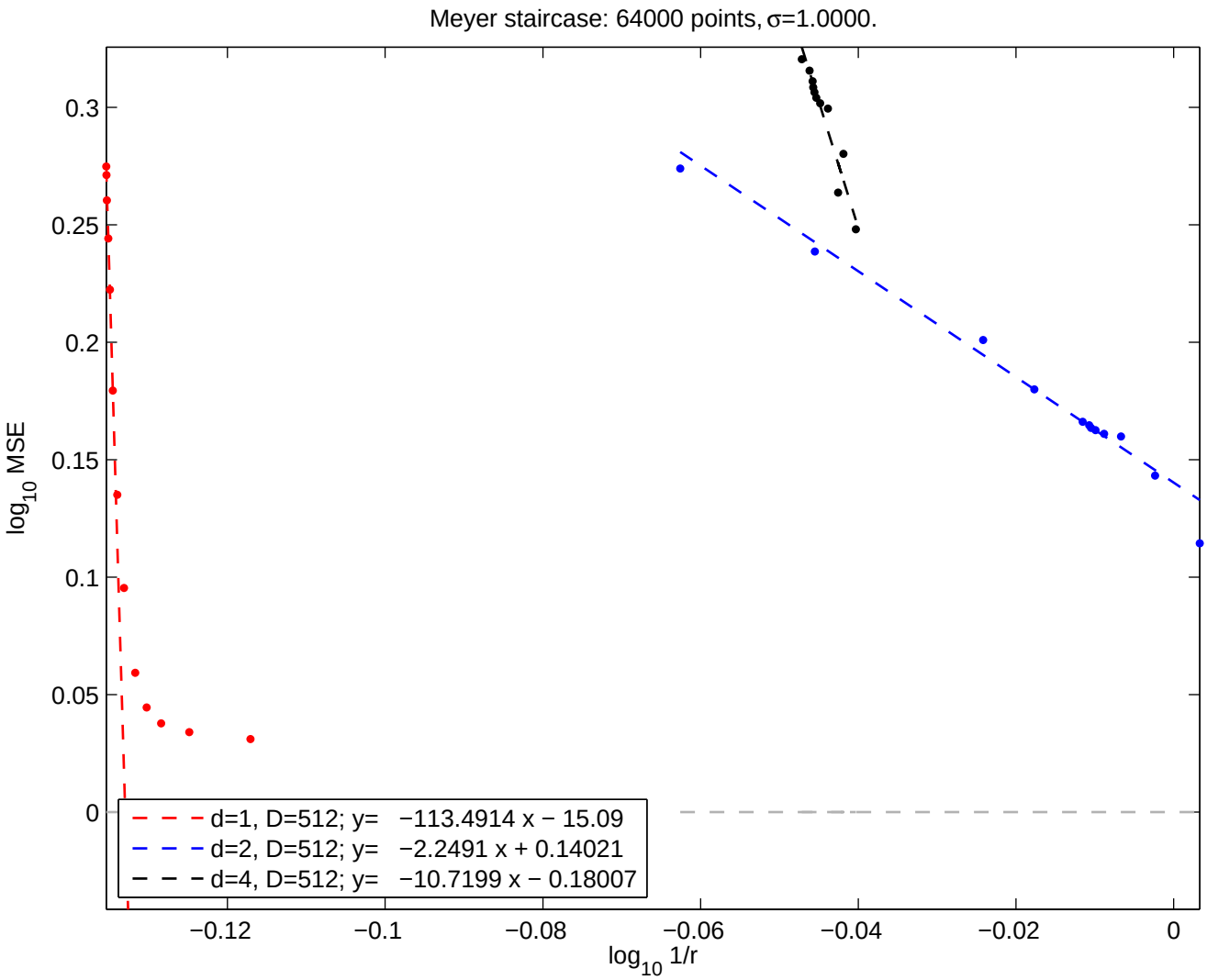

d−sphere Unif Noise: 16000 points, σ=0.1000.
d=1, D=100; y= −89.5919 x − 0.3842
d=2, D=100; y= −2.5749 x − 1.109
d=4, D=100; y= −4.4714 x − 0.78729
d=6, D=100; y= −4.443 x − 0.86476
d=8, D=100; y= −11.7805 x − 0.33441
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


d−sphere Unif Noise: 32000 points, σ=0.1000.
d=1, D=100; y= −122.4129 x − 0.502
d=2, D=100; y= −2.6495 x − 1.098
d=4, D=100; y= −4.1038 x − 0.88337
d=6, D=100; y= −4.6912 x − 0.87896
d=8, D=100; y= −5.2675 x − 0.84841
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
0
0.2
0.4
0.6
0.8
1


d−sphere Unif Noise: 64000 points, σ=0.1000.
d=1, D=100; y= −266.6304 x − 0.39922
d=2, D=100; y= −2.3992 x − 1.19
d=4, D=100; y= −4.4395 x − 0.8339
d=6, D=100; y= −4.7668 x − 0.89389
d=8, D=100; y= −4.4258 x − 0.93854
log10 MSE
log10 1/r
−0.5
−1
−1.5
−2
−2.5
−3
−3.5
0
0.2
0.4
0.6
0.8
1


Meyer staircase: 16000 points, σ=1.0000.
log10 MSE
log10 1/r
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.12
−0.1
−0.08
−0.06
−0.04
−0.02
0
d=1, D=512; y= −80.8785 x − 10.6738
d=2, D=512; y= −0.7309 x + 0.16299
d=4, D=512; y= −9.4112 x − 0.11292


Meyer staircase: 32000 points, σ=1.0000.
log10 MSE
log10 1/r
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.12
−0.1
−0.08
−0.06
−0.04
−0.02
0
d=1, D=512; y= −86.016 x − 11.3834
d=2, D=512; y= −1.4624 x + 0.15132
d=4, D=512; y= −9.5181 x − 0.12032


Meyer staircase: 64000 points, σ=1.0000.
log10 MSE
log10 1/r
0.3
0.25
0.2
0.15
0.1
0.05
0
−0.12
−0.1
−0.08
−0.06
−0.04
−0.02
0
d=1, D=512; y= −113.4914 x − 15.09
d=2, D=512; y= −2.2491 x + 0.14021
d=4, D=512; y= −10.7199 x − 0.18007