Entropy and Learning of Lipschitz Functions under Log-Concave Measures

Pierre Bizeul[1] and Boaz Klartag[2]

Abstract

We study regression of 1-Lipschitz functions under a log-concave measure μ on $\mathbb{R}^d$. We focus on the high-dimensional regime where the sample size n is subexponential in d, in which distribution-free estimators are ineffective. We analyze two polynomial-based procedures: the projection estimator, which relies on knowledge of an orthogonal polynomial basis of μ, and the least-squares estimator over low-degree polynomials, which requires no knowledge of μ whatsoever. Their risk is governed by the rate of polynomial approximation of Lipschitz functions in $L^2(\mu)$. When this rate matches the Gaussian one, we show that both estimators achieve minimax bounds over a wide range of parameters. A key ingredient is sharp entropy estimates for the class of 1-Lipschitz functions in $L^2(\mu)$, which are new even in the Gaussian setting.

1 Introduction

In this paper, we study the following regression problem. Given an unknown 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, we observe data

$$((X_1, Y_1), \ldots, (X_n, Y_n)),$$

where:

- The vectors $X_1, \ldots, X_n \in \mathbb{R}^d$ are independent random vectors that are distributed according to some Borel probability measure μ on $\mathbb{R}^d$ that may or may not be known to us.

- The numbers $Y_1, \ldots, Y_n \in \mathbb{R}$ are noisy observations of the function f evaluated at X_i, that is,

$$Y_i = f(X_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{1}$$

  where, throughout the paper, $\xi_1, \ldots, \xi_n$ are independent, real-valued Gaussian random variables of mean zero and variance σ^2, for some parameter $\sigma > 0$.

Our goal is to construct an estimator $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ of the function f, whose performance is measured by the $L^2(\mu)$-risk, defined via

$$\mathcal{R}(\hat{f}, f) := \mathbb{E}\|f - \hat{f}\|^2_{L^2(\mu)}. \tag{2}$$

There are various types of probability measures μ for which our analysis applies. We first consider a relatively simple case:

1.1 The Gaussian case

Consider first the case where $\mu = \gamma = \gamma_d$, the standard Gaussian measure on $\mathbb{R}^d$. A well-known fact (recalled below) is that any 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ can be approximated by polynomials in Gaussian space. Namely, for any $m \geq 1$, there exists a polynomial $P_m : \mathbb{R}^d \to \mathbb{R}$ of total degree at most m such that

$$\|f - P_m\|^2_{L^2(\gamma)} \leq \frac{1}{m+1}. \tag{3}$$

Here and throughout the paper, the degree of a multivariate polynomial refers to its total degree. More precisely, for a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ and the corresponding monomial

$$P(x) = \prod_{i=1}^d x_i^{\alpha_i} \qquad\qquad x = (x_1, \ldots, x_d) \in \mathbb{R}^d,$$

we define

$$\deg(P) := \sum_{i=1}^d \alpha_i =: |\alpha|.$$

Here $\mathbb{N} = \{0, 1, 2, \ldots\}$ stands for the set of all non-negative integers. The degree of a multivariate polynomial is the maximum of the degrees of its monomials. Note that the polynomial P_m in (3) is simply the orthogonal projection of f onto the finite-dimensional space of polynomials on $\mathbb{R}^d$ of degree at most m, denoted by $\mathcal{P}_{d,m}$. In particular, denoting by

$$(H_\alpha)_{\alpha \in \mathbb{N}^d}$$

the Hermite basis of orthogonal polynomials for γ, one can write

$$P_m = \sum_{\alpha \in \mathbb{N}^d, |\alpha| \leq m} \langle f, H_\alpha \rangle_{L^2(\gamma)} H_\alpha. \tag{4}$$

Our goal is to construct an estimator for the function f. Thanks to the polynomial approximation property (3), a natural idea is to estimate the polynomial $P_m \in \mathcal{P}_{d,m}$, for a suitable choice of degree m depending on n, d and σ. This reduces the nonparametric problem (1) to a parametric one. In view of (4), for a well-chosen m, one may construct an estimator $\hat{f}$ by empirically estimating the coefficients

$$f_\alpha := \langle f, H_\alpha \rangle_{L^2(\gamma)}.$$

Namely, we define

$$\hat{f} := \sum_{\alpha \in \mathbb{N}^d, |\alpha| \leq m} \hat{f}_\alpha H_\alpha, \tag{5}$$

where the coefficients $(\hat{f}_\alpha)_{|\alpha| \leq m}$ are defined as follows:

- First, for $\alpha = 0$, we estimate the Gaussian integral of f (its "barycenter")

$$a := f_0 = \int_{\mathbb{R}^d} f d\gamma$$

  via

$$\hat{a} := \hat{f}_0 = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} \sum_{i=1}^n f(X_i) + \frac{1}{n} \sum_{i=1}^n \xi_i. \tag{6}$$

  Clearly $\hat{a}$ is an unbiased estimator of a.

- Next, for any $\alpha \in \mathbb{N}^d$ with $|\alpha| > 0$ we define

$$\hat{f}_\alpha = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{a}) H_\alpha(X_i) \tag{7}$$

$$= \frac{1}{n} \sum_{i=1}^n (f(X_i) - a) H_\alpha(X_i) + \frac{1}{n} \sum_{i=1}^n \xi_i H_\alpha(X_i) + \frac{1}{n} \sum_{i=1}^n (\hat{a} - a) H_\alpha(X_i), \tag{8}$$

which is a biased estimator of f_α.

Note that the naïve unbiased estimator of f_α, namely

$$\check{f}_\alpha = \frac{1}{n} \sum_{i=1}^n Y_i H_\alpha(X_i), \tag{9}$$

may have an arbitrarily large variance, since we make no assumptions on the barycenter of f. If one assumes that the barycenter of f lies in some ball of fixed radius, independent of the dimension d and of the sample size n, then it makes sense to use the simpler estimator $\check{f}$ in place of $\hat{f}$.

Up to this minor variance reduction procedure, the estimator $\hat{f}$ is simply the projection estimator of f in the orthonormal basis of Hermite polynomials $(H_\alpha)_{\alpha \in \mathbb{N}^d}$.

1.2 The log-concave case

Moving away from the Gaussian setting, we aim to generalize the learning procedure from Section 1.1 to other measures. We shall assume that:

- The probability measure μ is a log-concave measure on $\mathbb{R}^d$, meaning that

$$d\mu(x) = e^{-V(x)} \, dx$$

for some convex potential $V : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$;

- The probability measure μ satisfies a polynomial approximation property: for any 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ and an integer $m \geq 1$, there exists a polynomial $P_m : \mathbb{R}^d \to \mathbb{R}$ of degree at most m such that

$$\|f - P_m\|_{L^2(\mu)}^2 \leq \Psi_\mu^2(m), \tag{10}$$

for some function $\Psi_\mu : \mathbb{N} \to \mathbb{R}^+$ decreasing to 0 as $m \to \infty$;

- for normalization purposes, let us assume that

$$\Psi_\mu(0) = 1. \tag{11}$$

In other words, for any 1-Lipschitz function f,

$$\mathrm{Var}_\mu(f) = \|f - \mathbb{E}_\mu f\|_{L^2(\mu)}^2 \leq \Psi_\mu^2(0) = 1.$$

A probability measure μ on $\mathbb{R}^d$ with finite second moments is *isotropic* if $\int_{\mathbb{R}^d} x_i d\mu(x) = 0$ for all i, and $\mathrm{Cov}(\mu) = \mathrm{Id}$, where $\mathrm{Cov}(\mu) = (\mathrm{Cov}_{ij}(\mu))_{i,j=1,\ldots,n} \in \mathbb{R}^{n \times n}$ is the covariance matrix, defined via

$$\mathrm{Cov}_{ij}(\mu) = \int_{\mathbb{R}^d} x_i x_j d\mu(x) - \int_{\mathbb{R}^d} x_i d\mu(x) \int_{\mathbb{R}^d} x_j d\mu(x).$$

Below we will mostly work with the isotropic normalization. The *projection estimator* is defined as follows:

Definition 1.1. Let μ be an isotropic log-concave measure on $\mathbb{R}^d$. Let $(P_\alpha)_{\alpha \in \mathbb{N}^d}$ be an orthonormal basis of polynomials in $L^2(\mu)$ with $\deg(P_\alpha) = |\alpha|$ for all α. Given observations of the form (1) and some parameter $m \in \mathbb{N}$, we define the projection estimator by

$$\hat{f} := \sum_{\alpha, \deg(P_\alpha) \leq m} \hat{f}_\alpha P_\alpha, \tag{12}$$

where

$$\hat{a} := \hat{f}_0 = \frac{1}{n} \sum_{i=1}^n Y_i, \tag{13}$$

and for all the other coefficients

$$\hat{f}_\alpha = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{a}) P_\alpha(X_i). \tag{14}$$

Let us mention that the Kannan–Lovász–Simonovits (KLS) conjecture suggests that the normalization (11) is equivalent to normalizing the largest variance over all directions:

$$c \leq \|\mathrm{Cov}(\mu)\|_{op} \leq 1 \tag{15}$$

for some universal constant $c > 0$, where $\| \cdot \|_{op}$ is the operator norm. For two functions a and b, we write $a \lesssim b$ or $a = O(b)$ if there exists a universal constant $C > 0$ such that $a \leq Cb$. We write $a \simeq b$ if $a \lesssim b$ and $b \lesssim a$. Using the best current bounds on the KLS constant [Kla23], one can take $c = c_n \simeq 1/\log n$ in (15).

Log-concave measures provide a natural generalization of the Gaussian case for two reasons. First, the behavior of Lipschitz and polynomial functions of a log-concave random vector is relatively well-understood. Second, although few explicit bounds are known, the polynomial approximation property (10) always holds—albeit possibly with a slowly decaying function Ψ_μ. A detailed discussion of these facts is provided in Section 2.

We prove the following upper bound on the performance of the *projection estimator*.

Theorem 1.2. *Let $n, d \geq 2$, and assume that the variance of the noise σ^2 satisfies*

$$\sigma^2 \leq d.$$

Define

$$m_0 = \lfloor \tfrac{\log n}{\log d} \rfloor.$$

We distinguish between two regimes:

- *If $d^5 \leq n \leq e^{\sqrt{d} \log d}$, set*

$$m := m_0 - 4.$$

For this choice of degree m we obtain the bound

$$\mathbb{E}\|f - \hat{f}\|_{L^2(\mu)}^2 \leq \Psi_\mu^2(m) + O\big(\tfrac{1}{d}\big). \tag{16}$$

- *If $e^{\sqrt{d} \log d} \leq n \leq e^{d \log d / 2}$, set*

$$m = m_0 - \left\lceil \frac{4 \log m_0}{\log(d/m_0)} \right\rceil.$$

For this choice of degree m we obtain the bound

$$\mathbb{E}\|f - \hat{f}\|_{L^2(\mu)}^2 \leq \Psi_\mu^2(m) + O\big(\tfrac{1}{m^2}\big). \tag{17}$$

4

The computation of the projection estimator requires apriori knowledge of an orthonormal basis of polynomials for μ. In the more general setting where μ is an *unknown* log-concave probability measure, one may instead use the polynomial that minimizes the empirical least-squares error.

Definition 1.3. Let μ be an isotropic log-concave measure. Given observations of the form (1) and some parameter $m \in \mathbb{N}$, we define the *least-squares estimator* by

$$\hat{f}_{\mathrm{LS}} := \underset{\deg(P) \leq m}{\operatorname{argmin}} \sum_{i=1}^{n} (P(X_i) - Y_i)^2, \tag{18}$$

That is, the sum on the right-hand side of (18) is a quadratic function on the finite-dimensional space $\mathcal{P}_{d,m}$ of polynomials of degree at most m on $\mathbb{R}^d$, and we define the estimator $\hat{f}_{\mathrm{LS}}$ to be any minimizer of this quadratic function. Note that the computation of the *least-squares estimator* requires no knowledge about the underlying measure μ.

We show that the performance of the least-squares estimator $\hat{f}_{\mathrm{LS}}$ is comparable to that of the projection estimator $\hat{f}$ in certain regimes.

Theorem 1.4. *Let $n, d \geq 2$, and assume that the variance of the noise σ^2 satisfies*

$$\sigma^2 \leq d.$$

Define

$$m_0 = \lfloor \tfrac{\log n}{\log d} \rfloor.$$

There exist universal constants $c_0, C_0 > 0$ such that the following hold:

- *If*

$$d^5 \leq n \leq e^{c_0 \log^2 d / \log \log d},$$

  set $m = m_0 - 4$. For this choice of degree m we have the bound,

$$\mathbb{E}\|f - f_{LS}\|_{L^2(\mu)}^2 \leq \Psi_\mu^2(m) + O\left(\tfrac{1}{d}\right). \tag{19}$$

- *If*

$$e^{c_0 \log^2 d / \log \log d} \leq n \leq e^{\frac{d^\beta}{C_0}},$$

  for some $\beta < 1/2$, define

$$\alpha = \frac{\log(C_0 \log n)}{\log d} < \tfrac{1}{2}, \qquad m = m_0 - 4 - \lfloor 2\alpha m_0 \rfloor.$$

  For this choice of degree m, assuming that $d \geq d(\beta)$ so that $m \geq 0$,

$$\mathbb{E}\|f - f_{LS}\|_{L^2(\mu)}^2 \leq \Psi_\mu^2(m) + O\left(\tfrac{1}{d}\right). \tag{20}$$

We also provide here lower bounds for the minimax rate of the learning problem (1). For a fixed probability measure μ on $\mathbb{R}^d$, define the minimax rate

$$\mathcal{R}_{n,d}^* = \inf_{\tilde{f}} \sup_{f} \mathcal{R}(f, \tilde{f}), \tag{21}$$

where the infimum runs over all estimators $\tilde{f}$ (i.e., all measurable functions of the data $(X_i, Y_i)_{i=1}^n$) and the supremum runs over all 1-Lipschitz functions f. A standard information-theoretic way of

providing a lower bound on $\mathcal{R}_{n,d}^*$ is the Fano method [Wai19], which requires entropy estimates. More precisely, let

$$d(f, g) := d_\mu(f, g) = \|f - g\|_{L^2(\mu)},$$

and let

$$B_{Lip} = \left\{ f : \mathbb{R}^d \to \mathbb{R} \ : \ f \text{ is 1-Lipschitz with } \int f^2 d\mu \leq 1 \right\}$$

be the unit ball of 1-Lipschitz functions for this metric. For $0 < \varepsilon < 1$, define

$$\mathcal{N}(B_{Lip}, \varepsilon, d_\mu)$$

to be the maximal size of an ε-separated set in B_{Lip} with respect to the metric $d = d_\mu$, and set

$$H_L^\mu(\varepsilon) = \log \mathcal{N}(B_{Lip}, \varepsilon, d_\mu),$$

the entropy of Lipschitz functions with respect to d_μ.

We lower bound H_L^μ when μ is an isotropic log-concave measure, with an improvement when it has a product structure. We say that a probability measure μ on $\mathbb{R}^d$ is a product measure if $X_1, \ldots, X_d$ are independent random variables whenever $X = (X_1, \ldots, X_d)$ has law μ.

Theorem 1.5. *Let μ be an isotropic log-concave measure on $\mathbb{R}^d$. Then for any ε with*

$$d^{-\eta} < \varepsilon < 1,$$

we have

$$\binom{d}{\lfloor c/\varepsilon \rfloor^2} \lesssim H_L^\mu(\varepsilon), \tag{22}$$

where $\eta < 1/4$ and $c > 0$ are universal constants. Moreover, if additionally μ is a product measure, then (22) holds with $\eta = 1/4$, that is, it holds in the range

$$d^{-1/4} < \varepsilon < 1.$$

As we will see in Section 4, the estimate (22) is tight up to the value of the constant c. Note that it is more conventional to define entropy via covering numbers rather than packing numbers. Since the two definitions are equivalent up to a factor of 2 in ε, this choice does not affect the result. Note that it is more conventional to define entropy via covering numbers rather than packing numbers. Since the two definitions are equivalent up to a factor of 2 in ε, this choice does not affect the result.

To the best of our knowledge, this result is new even in the Gaussian setting, and might be of independent interest. It allows us to derive minimax lower bounds for the learning problem (10).

Corollary 1.6. *Let μ be an isotropic log-concave measure on $\mathbb{R}^d$. Assume that the noise satisfies*

$$n^{-\kappa} \leq \sigma^2 \leq n$$

for some constant $\kappa > 0$. There exists a universal constant $c > 0$ such that if

$$n \leq e^{\frac{cd^{2\eta} \log d}{\kappa}},$$

the minimax risk is lower bounded as

$$\mathcal{R}_{n,d}^* \gtrsim (1 + \kappa) \frac{\log n}{\log d}. \tag{23}$$

Moreover, if additionally μ is a product measure, then the lower bound (23) holds in the range

$$n \leq e^{\frac{c\sqrt{d} \log d}{\kappa}}.$$

Thus, in the Gaussian case, or more generally, when μ is an isotropic log-concave measure satisfying

$$\Psi_\mu^2(m) \lesssim \frac{1}{m}, \tag{24}$$

we obtain matching bounds in certain regimes for both the projection and least-squares estimators. Specializing the previous bounds to the case where, say, $\kappa = 10$ we obtain the following:

Corollary 1.7. *Let $n, d \geq 2$, and let μ be an isotropic log-concave measure on $\mathbb{R}^d$ satisfying (24), such as the Gaussian measure or the uniform measure on the hypercube. Assume, for instance, that the noise parameter $\sigma > 0$ satisfies*

$$\frac{1}{n^{10}} \leq \sigma^2 \leq d.$$

Then the following hold:

- *The projection estimator and the least squares estimators achieves the minimax rate, up to a universal constant, in the range*

$$d^5 \leq n \leq e^{cd^{2\eta} \log d},$$

  where $c > 0$ is a universal constant. That is,

$$\frac{\log d}{\log n} \lesssim \mathcal{R}_{n,d}^* \leq \mathcal{R}(f, \hat{f}) \simeq \mathcal{R}(f, f_{LS}) \lesssim \frac{\log d}{\log n}. \tag{25}$$

- *If μ is additionally a product measure, then the projection estimator achieves minimax rate in the larger range*

$$d^5 \leq n \leq e^{c\sqrt{d} \log d}.$$

  For the least square estimator, there exists a universal constant $C > 0$ such that for any $0 < \beta < 1/2$, if

$$d^5 \leq n \leq e^{d^\beta/C},$$

  and $d \geq d(\beta)$ then the minimax rate is achieved up to a factor $(1 - 2\beta)^{-1}$:

$$\frac{\log d}{\log n} \lesssim \mathcal{R}_{n,d}^* \leq \mathcal{R}(f, f_{LS}) \lesssim (1 - 2\beta)^{-1} \frac{\log d}{\log n}.$$

In comparison, typical regression algorithms for smooth functions – such as nearest neighbors – require a number of samples that is at least exponential in the dimension. In contrast, our proposed algorithms attains the minimax rate in the high-dimensional regime, when the number of samples is merely subexponential in the dimension. As a concrete takeaway, consider learning a 1-Lipschitz function from noisy observations in $L^2(\gamma)$, where we recall that $\gamma = \gamma_d$ is the standard Gaussian measure in $\mathbb{R}^d$. In order to achieve accuracy up to a factor $\varepsilon > 0$, it suffices to use a sample size that grows only polynomially with the dimension:

$$n \simeq d^{\frac{c}{\varepsilon}}$$

for some constant $c > 0$. To the best of our knowledge, this result is new already in the Gaussian case. Our approach is related to the recent works of Eskenazis, Ivanishvili and Streck ([EI22], [EIS22]) on learning over the discrete hypercube, which rely on expansions in the orthonormal Walsh polynomial basis.

The remainder of this paper is organized as follows:

In Section 2, we review several properties of log-concave measures that will be used throughout the paper. We recall concentration inequalities for Lipschitz and polynomial functions, and present

polynomial approximation results for Lipschitz functions in $L^2(\mu)$. This background sets the stage for the statistical analysis.

In Section 3, we study in detail the two algorithms proposed for estimating Lipschitz functions: the projection estimator and the least-squares estimator. For both procedures we establish upper bounds on their $L^2(\mu)$ risk (Theorems 1.2 and 1.4).

In Section 4, we turn to lower bounds. We provide new estimates on the metric entropy of the class of 1-Lipschitz functions under isotropic log-concave measures (Theorem 1.5). As a consequence, we derive minimax lower bounds for the regression problem (1), showing that in certain regimes the upper and lower bounds match (Corollary 1.6).

2 Background on log-concave measures

In this section, we recall several properties of log-concave measures that are central to our study. We begin with the concentration properties of Lipschitz and polynomial functions under log-concave distributions. We then briefly review key facts about the Langevin semigroup associated with a log-concave measure. Finally, we discuss polynomial approximation of Lipschitz functions with respect to such measures.

2.1 Concentration of Lipschitz and polynomial functions

We recall that a probability measure μ on $\mathbb{R}^d$ is log-concave if it takes the form

$$\mu(dx) = e^{-V(x)} \, dx \qquad (26)$$

for some convex function $V : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. If the measure is supported in an affine subspace of $\mathbb{R}^d$, we require that its density relative to this affine subspace will take the form (26) for some convex function V. Gaussian measures, uniform distributions on convex bodies, and Dirac measures are all examples of log-concave probabilities. The convexity of V is known to imply strong concentration properties for μ.

We say that μ satisfies a Poincaré inequality with constant $C > 0$ if, for all locally Lipschitz functions f,

$$\mathrm{Var}_\mu(f) \le C \int |\nabla f|^2 \, d\mu. \qquad (27)$$

The best constant $C > 0$ in the Poincaré inequality is denoted by $C_P(\mu)$ and referred to as the *Poincaré constant* of μ. Namely,

$$C_P(\mu) := \sup_f \frac{\mathrm{Var}_\mu(f)}{\int |\nabla f|^2 \, d\mu},$$

where the supremum is taken over all locally Lipschitz, non-constant functions f. We also define

$$C_P^{\mathrm{Lip}}(\mu) := \sup_{f \in \mathrm{Lip}_1} \mathrm{Var}_\mu(f),$$

where the supremum is over all 1-Lipschitz functions f. Our normalization assumption (11) rewrites as

$$C_P^{\mathrm{Lip}}(\mu) = \Psi_\mu(0) = 1.$$

It is clear from the definitions that

$$C_P^{\mathrm{Lip}}(\mu) \le C_P(\mu).$$

However, a theorem of Emanuel Milman [Mil09] asserts that, for log-concave measures, these two quantities are equivalent up to a universal constant:

$$C_P(\mu) \simeq C_P^{\mathrm{Lip}}(\mu). \tag{28}$$

In other words, for log-concave measures, the Poincaré inequality is essentially saturated by Lipschitz functions. The KLS conjecture, originally formulated in [KLS95], proposes an even stronger statement: that the Poincaré inequality is actually saturated by linear functions. Namely, the trivial chain of inequalities

$$\|\mathrm{Cov}(\mu)\|_{\mathrm{op}} \leq C_P^{\mathrm{Lip}}(\mu) \leq C_P(\mu)$$

could be reversed up to universal constants. The best known estimate to date is due to [Kla23]:

$$C_P(\mu) \lesssim \log n. \tag{29}$$

A related but stronger functional inequality is the logarithmic Sobolev inequality. We say that μ satisfies a log-Sobolev inequality with constant $\rho > 0$ if, for all locally Lipschitz functions f,

$$\mathrm{Ent}_\mu(f^2) \leq 2\rho \int |\nabla f|^2 \, d\mu. \tag{30}$$

The best constant $\rho > 0$ for which this holds is denoted by $\rho_{LS}(\mu)$ and referred to as the *log-Sobolev constant* of μ. It always holds that

$$C_P(\mu) \leq \rho_{LS}(\mu),$$

and the log-Sobolev inequality is strictly stronger than the Poincaré inequality, as it implies sub-gaussian concentration rather than merely subexponential (see Proposition 2.1 below). In particular, unlike the Poincaré inequality, not all log-concave measures satisfy a log-Sobolev inequality. We refer to [Biz23] for further details. As a central example, the standard Gaussian measure satisfies

$$C_P(\gamma) = \rho_{LS}(\gamma) = 1.$$

As mentioned before, a Poincaré inequality implies exponential concentration for Lipschitz functions, while a log-Sobolev inequality implies stronger subgaussian concentration. These facts were observed by Gromov–Milman [GM83] (for Poincaré) and Herbst (for log-Sobolev), and are summarized in the following proposition.

Proposition 2.1. *Let μ be a probability measure on $\mathbb{R}^d$ and f a 1-Lipschitz function. There exists a universal constant $C > 0$ such that, for any $p \geq 1$,*

$$\|f\|_{L^p(\mu)} \leq Cp\sqrt{C_P(\mu)},$$

$$\|f\|_{L^p(\mu)} \leq C\sqrt{p}\sqrt{\rho_{LS}(\mu)}.$$

In particular, under our normalization (11), the moments of a Lipschitz function grow at most linearly with p. This fact can be reformulated in the context of Orlicz norms. A random variable X is said to be *sub-exponential* if there exists $K > 0$ such that

$$\mathbb{E}[\exp(|X|/K)] \leq 2,$$

and *sub-Gaussian* if

$$\mathbb{E}[\exp(X^2/K^2)] \leq 2.$$

The smallest such constant K defines the Orlicz norms $\|X\|_{\psi_1}$ and $\|X\|_{\psi_2}$ respectively. A well-known equivalent definition of the Orlicz norm is

$$\|X\|_{\psi_\alpha} \simeq \sup_{m \geq 1} \frac{\|X\|_m}{m^{1/\alpha}}.$$

Proposition 2.1 can be reformulated as :

$$\|f(X)\|_{\psi_1} \le C\sqrt{C_P(\mu)}, \quad \text{and} \quad \|f(X)\|_{\psi_2} \le C\sqrt{\rho_{LS}(\mu)}.$$

We recall the following standard Bernstein-type inequalities:

Proposition 2.2. *Let $X_1, \ldots, X_n$ be independent centered random variables. Then*

$$\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\right\|_{\psi_1} \lesssim \frac{1}{\sqrt{n}}\max_i \|X_i\|_{\psi_1}, \quad \left\|\frac{1}{n}\sum_{i=1}^{n} X_i\right\|_{\psi_2} \lesssim \frac{1}{\sqrt{n}}\max_i \|X_i\|_{\psi_2},$$

We refer to [Ver18] for background on subexponential and subgaussian distributions. As for polynomials, when the underlying measure is log-concave, we have the following estimates.

Proposition 2.3. *Let μ be a log-concave measure on $\mathbb{R}^d$, and let P be a degree-m polynomial. Abbreviate $\|P\|_q = \|P\|_{L^q(\mu)}$. Then there exists a universal constant $C > 0$ such that, for any $q \ge 2$,*

$$\|P\|_q \le \min\left(C^{(q-2)m}, (Cq)^m\right)\|P\|_2.$$

Moreover, for any $q \ge 1$, there exists $C_1 > 0$ such that

$$\|P\|_q \le (C_1 q)^m \|P\|_1.$$

Proof. The inequality
$$\|P\|_q \le (C_1 q)^m \|P\|_2 \le (C_2 q)^m \|P\|_1$$
was essentially established by Bourgain [Bou91], see also [NSV02]. It remains to interpolate for q close to 2. We may assume without loss of generality that $\|P\|_2 = 1$. By Hölder's inequality, for $2 \le q \le 4$,
$$\|P\|_q^q \le \|P\|_2^{4-q}\|P\|_4^{q-2} \le C^{m(q-2)},$$

which concludes the proof. $\qquad\square$

We remark that the following improvement holds when $\mu = \gamma$, the standard Gaussian measure (see [AS17, Proposition 5.48]):

Lemma 2.4. *Let P be a degree-m polynomial on $\mathbb{R}^d$. Then*

$$\|P\|_{L^q(\gamma)} \le (q-1)^{m/2}\|P\|_{L^2(\gamma)}.$$

We will also need a classical anti-concentration result for polynomials in log-concave variables, due to Carbery and Wright ([CW01, Theorem 8])

Theorem 2.5. *Let X be a log-concave random vector in $\mathbb{R}^d$, and let P be a polynomial of degree at most m such that $\mathbb{E}P^2(X) = 1$. Then for all $t > 0$,*

$$\mathbb{P}\left(|P(X)| \le t\right) \le Cmt^{1/m},$$

where $C > 0$ is a universal constant.

2.2 Langevin semigroup

We now briefly recall some basic facts about the semigroup associated with a log-concave measure. For a detailed exposition, we refer to [BGL13]. Let μ be a log-concave probability measure on $\mathbb{R}^d$ with density

$$\mu(dx) = e^{-V(x)} \, dx$$

for a convex $V : \mathbb{R}^d \to \mathbb{R}$. The probability measure μ is associated with the symmetric diffusion operator

$$L := \Delta - \nabla V \cdot \nabla, \tag{31}$$

which satisfies, for smooth, compactly-supported functions $f, g : \mathbb{R}^d \to \mathbb{R}$,

$$\int f \, Lg \, d\mu = \int g \, Lf \, d\mu = - \int \nabla f \cdot \nabla g \, d\mu. \tag{32}$$

Consider the Friedrich self-adjoint extension of the operator L to a self-adjoint operator on $L^2(\mu)$, which is also denoted by L. The corresponding semigroup is

$$T_t := e^{tL} \qquad\qquad (t \geq 0). \tag{33}$$

This semigroup is Markovian and admits an explicit probabilistic representation: if $(X_t)_{t \geq 0}$ solves the stochastic differential equation

$$dX_t = \sqrt{2} \, dB_t - \nabla V(X_t) \, dt,$$

where (B_t) is standard Brownian motion in $\mathbb{R}^d$, then (X_t) is a Markov process, and

$$T_t f(x) = \mathbb{E}[f(X_t) \mid X_0 = x].$$

It follows that T_t is a contraction on $L^p(\mu)$ for all $p \geq 1$. The semigroup T_t has been widely used to establish functional inequalities for μ, since it continuously interpolates between $T_0 f = f$ and $T_\infty f = \int f \, d\mu$. The rate at which $T_t f$ converges to the mean is governed by the gradient of f:

Lemma 2.6. *Let $f \in L^2(\mu)$ be a smooth function with $\int_{\mathbb{R}^d} |\nabla f|^2 d\mu < \infty$. Then*

$$\|T_t f\|_{L^2(\mu)}^2 \geq \|f\|_{L^2(\mu)}^2 - 2t \int |\nabla f|^2 \, d\mu.$$

Proof. The argument is standard, we sketch the computation:

$$\frac{d}{dt} \|T_t f\|_{L^2(\mu)}^2 = 2 \langle L T_t f, \, T_t f \rangle_{L^2(\mu)}$$

$$= -2 \int |\nabla T_t f|^2 \, d\mu$$

$$\geq -2 \int T_t |\nabla f|^2 \, d\mu$$

$$\geq -2 \int |\nabla f|^2 \, d\mu,$$

where we used the standard gradient bound that follows from log-concavity

$$|\nabla T_t f|^2 \leq T_t |\nabla f|^2,$$

and the fact that T_t is a contraction. $\qquad\square$

Since T_t acts as a local averaging operator, one may expect smoothing properties. It is well-known e.g. [KL25] that T_t maps bounded functions to Lipschitz functions:

Lemma 2.7. *For every bounded function f and any $t > 0$, we have*

$$\|T_t f\|_{\mathrm{Lip}} \leq \frac{\|f\|_\infty}{\sqrt{t}}.$$

2.3 Polynomial approximation of Lipschitz functions

2.3.1 Dimension 1

For a log-concave measure μ on $\mathbb{R}$, we define $\Psi_\mu(m)$ as the best function such that, for any 1-Lipschitz function f and integer $m \geq 1$,

$$E_m(\mu, f) := \inf_{\deg(P_m) \leq m} \|f - P_m\|_{L^2(\mu)} \leq \Psi_\mu(m). \tag{34}$$

We begin with a well-known qualitative result.

Proposition 2.8. *Let μ be a measure on $\mathbb{R}^d$ such that there exists $\varepsilon > 0$ such that for all $\theta \in B_2(0, \varepsilon)$,*

$$\int e^{\theta \cdot x} \, \mu(dx) < \infty.$$

Then polynomials are dense in $L^2(\mu)$. Moreover, the convergence is uniform over the class $\mathcal{F}_{\mathrm{Lip}}$ of 1-Lipschitz functions:

$$E_m(\mu, \mathcal{F}) := \sup_{f \in \mathcal{F}} E_m(f, \mu) \longrightarrow 0 \quad \text{as } m \to \infty.$$

Proof. Let $f \in L^2(\mu)$ be orthogonal to all polynomials, and define the signed measure $\mu_f(dx) = f(x)\mu(dx)$. By the Cauchy–Schwarz inequality, for θ small enough:

$$\left(\int e^{\theta \cdot x} \, \mu_f(dx) \right)^2 \leq \left(\int f^2 \, d\mu \right) \left(\int e^{2\theta \cdot x} \, \mu(dx) \right) < \infty.$$

Thus, μ_f admits a Laplace transform defined in a neighborhood of 0, and all of its derivatives at the origin vanish due to the orthogonality condition. It follows that $\mu_f = 0$, hence $f = 0$ in $L^2(\mu)$. The uniform convergence follows by compactness of the set of centered 1-Lipschitz functions in $L^2(\mu)$. $\square$

In particular, since any log-concave probability measure satisfies the exponential integrability condition, we have

$$\Psi_\mu(m) \longrightarrow 0$$

as $m \to \infty$.

We now turn to quantitative statements. A classical result in this direction is the quantitative Weierstrass approximation theorem, going back to Bernstein and to Jackson. It asserts that for a 1-Lipschitz function f on the interval $[-1, 1]$,

$$\inf_{\deg(P) \leq m} \|f - P\|_{L^\infty([-1,1])} \leq \frac{C}{m}. \tag{35}$$

Let us describe in some detail how to obtain an L^2 version of this result. Let μ be the uniform probability measure on $[-1, 1]$. The orthogonal polynomials with respect to μ are the Legendre polynomials,

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} \left(x^2 - 1 \right)^n,$$

which satisfy

$$\int_{-1}^{1} P_n(x) P_m(x) \, d\mu(x) = \frac{1}{2n + 1} \delta_{nm}.$$

They also satisfy the differential equation

$$\frac{d}{dx}\left((1-x^2)P_n'(x)\right) + n(n+1)P_n(x) = 0. \tag{36}$$

Integrating by parts gives

$$\int_{-1}^{1} P_n'(x)P_m'(x)(1-x^2)\,dx = \frac{2n(n+1)}{2n+1}\delta_{nm}. \tag{37}$$

Let us normalize the Legendre polynomials as

$$p_n := \frac{P_n}{\sqrt{2n+1}},$$

so that (p_n) forms an orthonormal basis in $L^2(\mu)$. Any function $f \in L^2(\mu)$ can be expanded as

$$f = \sum_{k\geq 0} \langle f, p_k\rangle_{L^2(\mu)} p_k = \sum_{k\geq 0} f_k p_k.$$

Observe that if $f' \in L^2((1-x^2)\mu)$, then

$$f' = \sum_{k\geq 1} f_k p_k',$$

and by orthogonality using (37), we obtain

$$\int (f')^2(1-x^2)\,d\mu = \sum_{k\geq 1} k(k+1) f_k^2. \tag{38}$$

In particular,

$$E_m^2(\mu, f) = \sum_{k\geq m+1} f_k^2 \leq \frac{1}{(m+1)(m+2)} \int (f')^2(1-x^2)\,d\mu. \tag{39}$$

Following Jackson's theorem, an extensive body of work was devoted to the study of polynomial approximation on $\mathbb{R}$ under more general probability measures (or "weights"). A good reference is the survey [Lub07]. Let us denote

$$\mu_\alpha := \frac{1}{Z_\alpha} e^{-|x|^\alpha},$$

where Z_α is a normalizing constant. It can be shown that polynomials are dense in $L^2(\mu_\alpha)$ if and only if $\alpha \geq 1$. When $\alpha > 1$, it was proved by Freud [Fre77] and Lubinsky and Levin [LL87] that, for sufficiently regular functions f,

$$E_m^2(f, \mu_\alpha) \lesssim m^{2-2/\alpha} \int_{\mathbb{R}} (f')^2\,d\mu_\alpha. \tag{40}$$

The case $\alpha = 1$ is different: it can be shown that the set

$$\left\{ f \in L^2(\mu_1) : \int f\,d\mu_1 = 0, \quad \int (f')^2\,d\mu_1 \leq 1 \right\}$$

is not compact in $L^2(\mu_1)$. We refer to [BGL13] for details. As a consequence, a bound of the form (40) cannot hold with any fixed rate. Nevertheless, a corollary of a result by [Lub07] shows that if f is 1-Lipschitz, then

$$E_m^2(f, \mu_1) \lesssim \frac{1}{\log^2(m+1)}. \tag{41}$$

As a consequence, we may state a universal approximation rate for log-concave probability measures on the real line:

13

Lemma 2.9. *Let μ be a log-concave probability measure on $\mathbb{R}$ with unit variance. Then for any 1-Lipschitz function f and $m \geq 1$,*

$$E_m^2(\mu, f) \lesssim \frac{1}{\log^2(1+m)}.$$

Proof. This follows from the fact that if ρ is a log-concave density on $\mathbb{R}$ with unit variance and barycenter at 0, then

$$\rho(x) \leq C \, e^{-|x|/C}$$

for some universal constant $C > 0$. A proof of this estimate can be found in [Bob03]. □

We also mention that, in sharp contrast with the two-sided exponential distribution, the one-sided exponential enjoys a much faster approximation rate. Denote by $\nu = e^{-x} \, \mathbb{1}_{\mathbb{R}+}(x) \, dx$. Then for 1-Lipschitz functions f,

$$E_m^2(\nu, f) \lesssim \frac{1}{m},$$

see e.g., [BK25].

2.3.2 Higher dimensions

In higher dimensions, quantitative results on polynomial approximation are scarce. A notable exception is the case of the Gaussian measure γ_d on $\mathbb{R}^d$, which admits an explicit orthogonal basis of polynomials: the Hermite polynomials. In dimension one, the n-th Hermite polynomial is defined via the Rodrigues formula:

$$H_n(x) := (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2}, \tag{42}$$

and satisfies the orthogonality relation:

$$\int_{-\infty}^{+\infty} H_n(x) H_m(x) \, e^{-x^2/2} \, dx = n! \sqrt{2\pi} \, \delta_{n,m}. \tag{43}$$

In dimension d, for a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d)$, define

$$H_\alpha(x) := H_{\alpha_1}(x_1) \cdots H_{\alpha_d}(x_d).$$

The Hermite polynomials form an orthogonal basis of $L^2(\gamma_d)$ and are also eigenfunctions of the differential operator

$$L := \Delta + x \cdot \nabla,$$

which is the infinitesimal generator of the Ornstein–Uhlenbeck semigroup:

$$T_t f(x) := \mathbb{E}\left[f\left(\delta_t x + \sqrt{1 - \delta_t^2} G \right) \right],$$

where $G \sim \gamma_d$ and $\delta_t = e^{-t}$.

Proposition 2.10. *For any multi-index α, we have:*

$$L H_\alpha = -|\alpha| \, H_\alpha,$$

and consequently,

$$T_t H_\alpha = e^{-t|\alpha|} H_\alpha.$$

Given a function $f \in L^2(\gamma_d)$, we may decompose it in the Hermite basis as

$$f = \sum_\alpha f_\alpha H_\alpha.$$

Using the integration by parts formula (32), we compute the gradient energy:

$$\int \|\nabla f\|^2 \, d\gamma_d = \int (-Lf) f \, d\gamma_d$$
$$= \sum_\alpha |\alpha| f_\alpha^2.$$

Finally,

$$E_m^2(\gamma_d, f) = \sum_{|\alpha| \geq m+1} f_\alpha^2 \leq \frac{1}{m+1} \sum_\alpha |\alpha| f_\alpha^2$$
$$\leq \frac{1}{m+1} \int |\nabla f|^2 \, d\gamma_d.$$

This is an instance of the tensorization principle established in [BK25].

Proposition 2.11 (Tensorization). *Let μ_i be a probability measure on $\mathbb{R}$ for $i = 1, \ldots, d$. Assume that for all i,*

$$\sum_{\alpha \geq 1} \varphi_i(\alpha) f_\alpha^2 \leq \int_{\mathbb{R}} (f')^2 w_i(x) \, d\mu_i,$$

for some positive functions $(\varphi_i)_{1 \leq i \leq d}$ and $(w_i)_{1 \leq i \leq d}$, where f_α denotes the coefficients in the orthonormal polynomial basis of μ_i. Let $\mu := \mu_1 \otimes \cdots \otimes \mu_d$. Then for all smooth $f \in L^2(\mu)$,

$$\sum_{|\alpha| \geq 1} \varphi(\alpha) f_\alpha^2 \leq \int_{\mathbb{R}^d} \sum_{i=1}^d w_i(x_i)(\partial_i f)^2 \, d\mu,$$

where $\varphi(\alpha) := \sum_i \varphi_i(\alpha_i)$ and we set $\varphi_i(0) := 0$, $|\alpha| := \sum_i \alpha_i$. In particular, defining

$$\Phi(m) := \inf_{|\alpha| = m} \varphi(\alpha),$$

we obtain the approximation bound

$$E_m^2(\mu, f) \leq \frac{1}{\Phi(m+1)} \int_{\mathbb{R}^d} \sum_{i=1}^d w_i(x_i)(\partial_i f)^2 \, d\mu.$$

Moreover, if f is 1-Lipschitz,

$$E_m^2(\mu, f) \leq \frac{1}{\Phi(m+1)} \mathbb{E}\left[\max_i w_i(X_i)\right].$$

Let us illustrate Proposition 2.11 in concrete examples. For $1 < \beta \leq 2$, define the product measure

$$\mu_\beta^{\otimes d} := \mu_\beta \otimes \cdots \otimes \mu_\beta.$$

Recall that in dimension 1, we have

$$E_m^2(\mu_\beta, f) \lesssim m^{2/\beta - 2} \int_{\mathbb{R}} (f')^2 \, d\mu_\beta,$$

15

i.e., the tail bound

$$\sum_{k \geq m+1} f_k^2 \lesssim m^{2/\beta-2} \int_{\mathbb{R}} (f')^2 \, d\mu_\beta.$$

Using summation by parts, we deduce that for any $0 < \delta \leq 1$,

$$\sum_{k \geq 1} \frac{k^{2-2/\beta} f_k^2}{\log^{1+\delta}(1+k)} \leq \frac{C}{\delta} \int_{\mathbb{R}} (f')^2 \, d\mu_\beta,$$

for some constant $C > 0$. Define

$$\varphi(x) := \frac{x^{2-2/\beta}}{\log^{1+\delta}(x)}.$$

Since $\varphi(x)/x$ decreases on $(1, \infty)$, we obtain

$$\Phi(m) := \inf_{|\alpha|=m} \sum_i \varphi(\alpha_i) = \varphi(m).$$

By tensorization, this yields

$$E_m^2(\mu_\beta^{\otimes d}, f) \lesssim \frac{\log^{1+\delta}(m)}{\delta \, m^{2-2/\beta}} \int_{\mathbb{R}^d} \|\nabla f\|^2 \, d\mu_\beta^{\otimes d}.$$

Choosing

$$\delta^* := \max\left(1, \frac{1}{\log\log m}\right),$$

we obtain

$$E_m^2(\mu_\beta^{\otimes d}, f) \lesssim \frac{\log m \log\log m}{m^{2-2/\beta}} \int_{\mathbb{R}^d} \|\nabla f\|^2 \, d\mu_\beta^{\otimes d}. \tag{44}$$

For the case $\beta = 1$, it was proved in [BK25] that

$$\sum_{k=1}^{\infty} \log^2(e+k) f_k^2 \lesssim \int_{\mathbb{R}} \log^2(e+|x|)(f')^2 \, d\mu_1. \tag{45}$$

Hence, tensorization gives

$$E_m^2(f, \mu_1^{\otimes d}) \lesssim \frac{1}{\log^2(1+m)} \int_{\mathbb{R}^d} \sum_{i=1}^{d} \log^2(e+|x_i|)(\partial_i f)^2 \, d\mu_1^{\otimes d}.$$

If f is 1-Lipschitz, then

$$E_m^2(f, \mu_1^{\otimes d}) \lesssim \frac{\mathbb{E}\left[\max_i \log^2(e+|X_i|)\right]}{\log^2(m)} \lesssim \frac{\log^2 \log d}{\log^2 m}.$$

In contrast, for the product of one-sided exponential measures, the approximation rate is much better. Let $\nu^{\otimes d}$ be the d-fold product of the one-sided exponential distribution. Then for 1-Lipschitz f,

$$E_m^2(\nu^{\otimes d}, f) \lesssim \frac{\log d}{m},$$

see [BK25].

We conclude this section with an interesting dimensional effect of tensorization, which lies at the core of the entropy estimates of Section 4. Let μ be the uniform probability measure on $[-1, 1]$. Recall that for sufficiently regular f, using (38),

$$\sum_{k \geq 1} k^2 f_k^2 \leq \int (f')^2 (1-x^2) \, d\mu \leq \int (f')^2 \, d\mu.$$

16

The rate is quadratic, much faster than the linear rate observed for the Gaussian measure, for example. For the uniform measure on the hypercube $\mu^{\otimes d}$, the tensorization principle yields, in particular,

$$E_m^2(\mu, f) \leq \frac{1}{\Phi(m+1)} \int \|\nabla f\|^2 \, d\mu^{\otimes d},$$

where

$$\Phi(m) = \inf_{|\alpha| = m} \sum_i \alpha_i^2.$$

The key difference is that here the function $\varphi(x) = x^2$ is convex, so that $\varphi(x)/x$ is increasing (as opposed to decreasing). Therefore,

$$\Phi(m) = m, \qquad \text{for } m \leq d,$$

while for $m \geq d$,

$$\Phi(m) \simeq \frac{m^2}{d}.$$

The takeaway is that when the degree m is smaller than or comparable to the dimension d, the rate cannot be better than the Gaussian one, i.e., linear in m.

More precisely, let μ be an isotropic product measure on $\mathbb{R}^d$, and let Φ_μ denote the best function such that for all sufficiently regular functions f,

$$E_m^2(\mu, f) \leq \frac{1}{\Phi_\mu(m)}.$$

Then, for $m \leq d$,

$$\Phi_\mu(m) \leq \Phi_\gamma(m) = m + 1.$$

The corresponding extremal function f is the multilinear polynomial of degree $m \leq d$ defined by

$$f(x) = \prod_{i=1}^m x_i.$$

Whenever the measure μ is isotropic and of product form, the function f belongs to the tensor basis of orthonormal polynomials. Therefore, for all $k < m$,

$$E_k^2(\mu, f) = \|f - 0\|_{L^2(\mu)}^2 = 1.$$

On the other hand, a direct computation shows that

$$\int \|\nabla f\|^2 \, d\mu = m.$$

Hence,

$$\Phi_\mu(m-1) \leq \frac{\int \|\nabla f\|^2 \, d\mu}{E_{m-1}^2(\mu, f)} \leq m.$$

We will see in Section 4 that if μ is additionally log-concave, this remains true when restricting to Lipschitz functions, at least for $m \leq \sqrt{d}$.

3 Empirical computation of an approximating polynomial

In this section we analyze in detail the empirical procedures introduced in the introduction. Given observations (1), our goal is to construct a polynomial estimator of f, taking advantage of the approximation property (10). We focus on two natural algorithms: the projection estimator, which relies on an orthogonal polynomial basis of μ, and the least-squares estimator, which requires no structural knowledge of μ.

3.1 The projection estimator

We fix a 1-Lipschitz function f, and assume that μ is a log-concave probability measure on $\mathbb{R}^d$ with polynomial approximation rate:

$$\sup_f \inf_{\deg(P) \le m} \|f - P\|_{L^2(\mu)} \le \Psi_\mu(m),$$

where the sup runs over all 1-Lipschitz functions f. Recall that for normalization purposes, we assume that $\Psi_\mu(0) = 1$, which amounts to the bound $C_P(\mu) \lesssim 1$.

We decompose the function f as

$$f = f - \mathbb{E}_\mu f + \mathbb{E}_\mu f =: \tilde{f} + a, \tag{46}$$

where $a = \mathbb{E}_\mu f$ is a constant and $\tilde{f}$ is mean-zero.

Recall that we denote by $X_1, \ldots, X_n$ the observed i.i.d. samples from μ. The integer n denotes the sample size used in the algorithm, while m denotes the maximal polynomial degree used. Finally, we define

$$D = D(d, m) := \binom{d + m}{m},$$

which is the dimension of the space $\mathcal{P}_{d,m}$ of multivariate polynomials of total degree at most m in $\mathbb{R}^d$. Let us further denote by $(p_k)_{0 \le k \le D-1}$ an orthonormal basis of $\mathcal{P}_{d,m} \subseteq L^2(\mu)$. Thus, we may decompose f as

$$f = \sum_{k=0}^{D-1} f_k \, p_k + f_{>m} = a + \sum_{k=1}^{D-1} f_k \, p_k + f_{>m}, \tag{47}$$

where for all $1 \le k \le D-1$,

$$f_k := \langle f, p_k \rangle_{L^2(\mu)} = \langle \tilde{f}, p_k \rangle_{L^2(\mu)},$$

and

$$\|f_{>m}\|_{L^2(\mu)} \le \Psi(m).$$

Recall that the empirical estimator of the mean is given by

$$\hat{a} := \hat{f}_0 = \frac{1}{n} \sum_{i=1}^{n} Y_i,$$

and for $1 \le k \le D-1$, we define the empirical coefficients as

$$\begin{aligned} \hat{f}_k &= \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{a}) \, p_k(X_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} (Y_i - a) \, p_k(X_i) + \frac{1}{n} \sum_{i=1}^{n} (\hat{a} - a) \, p_k(X_i) \\ &=: \hat{f}_k^* + \delta_k. \end{aligned} \tag{48}$$

Here, $\hat{f}_k^*$ is the unbiased component of the estimator, satisfying

$$\mathbb{E}[\hat{f}_k^*] = f_k.$$

The projection estimator is then given by

$$\hat{f} := \hat{a} + \sum_{k=1}^{D-1} \hat{f}_k \, p_k.$$

Our goal is to prove the following result, previously stated in the introduction.

Theorem 3.1. *Under the above assumptions, the projection estimator satisfies*

$$\mathbb{E}\|f - \hat{f}\|^2_{L^2(\mu)} \leq \Psi^2_\mu(m) + \frac{(Cm^2 + 4\sigma^2)D}{n}, \tag{49}$$

for some absolute constant $C > 0$.

Furthermore, in the Gaussian setting, we have the sharper bound

$$\mathbb{E}\|f - \hat{f}\|^2_{L^2(\gamma)} \leq \frac{1}{m} + \frac{(Cm + 4\sigma^2)D}{n}, \tag{50}$$

for some absolute constant $C > 0$.

Proof. Let P_m denote the orthogonal projection of f onto the space $\mathcal{P}_{d,m}$ of polynomials of degree at most m. Then:

$$\mathbb{E}\|f - \hat{f}\|^2_2 = \mathbb{E}\|f - P_m\|^2_2 + \mathbb{E}\|\hat{f} - P_m\|^2_2$$

$$\leq \Psi^2(m) + \text{Var}(\hat{a}) + \sum_{k=1}^{D-1} \mathbb{E}(\hat{f}_k - f_k)^2$$

$$\leq \Psi^2(m) + \text{Var}(\hat{a}) + 2\sum_{k=1}^{D-1} \mathbb{E}(\hat{f}_k^* - f_k)^2 + 2\sum_{k=1}^{D-1} \mathbb{E}\delta_k^2$$

$$= \Psi^2(m) + \frac{\text{Var}(Y_1)}{n} + 2\sum_{k=1}^{D-1} \text{Var}(\hat{f}_k^*) + 2\sum_{k=1}^{D-1} \mathbb{E}\delta_k^2$$

$$\leq \Psi^2(m) + \frac{1+\sigma^2}{n} + \frac{2}{n}\sum_{k=1}^{D-1} \text{Var}\left((Y_1 - a)p_k(X_1)\right) + 2\sum_{k=1}^{D-1} \mathbb{E}\delta_k^2, \tag{51}$$

where in the last passage we used that $Var_\mu(f) \leq \Psi_\mu(0) = 1$. We first bound the third term in (51). Let (X, Y) denote a copy of (X_1, Y_1). Observe that

$$Y - a = f(X) + \xi - a = \tilde{f}(X) + \xi,$$

where $\tilde{f} = f - a$ is centered. Then, for any $1 \leq k \leq D$,

$$\text{Var}\left((Y - a)p_k(X)\right) \leq \mathbb{E}\left((Y-a)^2 p_k^2(X)\right)$$

$$= \mathbb{E}\left(\tilde{f}^2(X)p_k^2(X)\right) + \mathbb{E}\left(\xi^2 p_k^2(X)\right)$$

$$= \mathbb{E}\left(\tilde{f}^2(X)p_k^2(X)\right) + \sigma^2.$$

Now we apply Hölder's inequality with exponents $q = m + 1$ and $q^* = 1 + 1/m$:

$$\mathbb{E}\left[\tilde{f}^2(X)p_k^2(X)\right] \leq \left(\mathbb{E}\tilde{f}^{2q}(X)\right)^{1/q}\left(\mathbb{E}p_k^{2q^*}(X)\right)^{1/q^*}$$

$$\leq C\|\tilde{f}(X)\|^2_{2m+2},$$

as follows from Proposition 2.3. Recalling that $C_P(\mu) \lesssim 1$, by Proposition 2.1, we have

$$\|\tilde{f}(X)\|^2_{2m+2} \lesssim m^2, \tag{52}$$

since $\tilde{f}$ is 1-Lipschitz and centered. Now we bound the fourth term in (51). Define

$$S_n := \frac{1}{n}\sum_{i=1}^{n} p_k(X_i).$$

19

Then:

$$\mathbb{E}\delta_k^2 = \mathbb{E}\left(\hat{a} - a\right)^2 S_n^2$$

$$= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \left(\tilde{f}(X_i) + \xi_i\right)\right)^2 S_n^2$$

$$= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \tilde{f}(X_i)\right)^2 S_n^2 + \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \xi_i\right)^2 S_n^2$$

$$= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \tilde{f}(X_i)\right)^2 S_n^2 + \frac{\sigma^2}{n^2}\mathbb{E}S_n^2.$$

Note that $\mathbb{E}S_n^2 = \mathbb{E}P_k^2(X)/n = 1/n$. By again using Hölder's inequality with $q = m + 1$, and bounding $\|S_n\|_{q^*} \le \|p_k(X)\|_{q^*}$:

$$\mathbb{E}\delta_k^2 \le \|\frac{1}{n}\sum_{i=1}^n \tilde{f}(X_i)\|_q^2 \cdot \|p_k(X)\|_{q^*}^2 + \frac{\sigma^2}{n^3}$$

$$\lesssim \|\frac{1}{n}\sum_{i=1}^n \tilde{f}(X_i)\|_{m+1}^2 + \frac{\sigma^2}{n^2}.$$

By Proposition 2.2, we know that for a 1-Lipschitz function:

$$\|\frac{1}{n}\sum_{i=1}^n \tilde{f}(X_i)\|_{\psi_1} \lesssim \frac{1}{\sqrt{n}}\|\tilde{f}(X)\|_{\psi_1} \lesssim \frac{1}{\sqrt{n}}.$$

Hence,

$$\|\frac{1}{n}\sum_{i=1}^n \tilde{f}(X_i)\|_{m+1}^2 \lesssim \frac{m^2}{n}. \tag{53}$$

Plugging everything back into (51), we obtain:

$$\mathbb{E}\|f - \hat{f}\|_2^2 \le \Psi_\mu^2(m) + \frac{Cm^2 D}{n} + \frac{4\sigma^2 D}{n},$$

for some absolute constant $C > 0$, as claimed. For the "Furthermore" part, we replace m^2 in (52) and in (53) by $m \cdot \min\{m, \rho_{LS}(X)\}$ which equals m in the Gaussian case.

3.1.1 Proof of Theorem 1.2

Let us explain how to deduce Theorem 1.2 from Theorem 3.1. Set

$$5 \le m_0 = \left\lfloor \frac{\log n}{\log d} \right\rfloor \le \frac{d}{2}$$

and observe that

$$\binom{d + m_0}{m_0} \le \left(\frac{e(d + m_0)}{m_0}\right)^{m_0}$$

$$\le \left(\frac{5d}{m_0}\right)^{m_0}$$

$$\le d^{m_0} \le n.$$

For any integer $1 \leq p \leq m_0$ we set

$$m_p = m_0 - p.$$

Recall that $D = D(d, m_p) = \binom{d + m_p}{m_p}$. According to the preceding inequality, we have :

$$\frac{\binom{d+m_p}{m_p}}{n} \leq \frac{\binom{d+m_0-p}{m_0-p}}{\binom{d+m_0}{m_0}} = \frac{(d+m_0-p)!}{(d+m_0)!} \cdot \frac{m_0!}{(m_0-p)!}$$

$$\leq \left(\frac{m_0}{d}\right)^p.$$

Plugging this into Theorem 3.1, we get for the choice of $m = m_p$,

$$\mathbb{E}\|f - \hat{f}\|^2_{L^2(\mu)} \leq \Psi^2_\mu(m_p) + \left(Cm_0^2 + 4\sigma^2\right)\left(\frac{m_0}{d}\right)^p.$$

We choose

$$p = \max\left(4, \left\lceil \frac{4\log m_0}{\log d/m_0} \right\rceil\right).$$

In the first regime,

$$n \leq e^{\sqrt{d}\log d}$$

and $p = 4$ while $m_0 \leq \sqrt{d}$. We have

$$\mathbb{E}\|f - \hat{f}\|^2_{L^2(\mu)} \leq \Psi^2_\mu(m_0 - 4) + C'd(1/\sqrt{d})^4$$

which is what we wanted to proved. In the second regime, we have

$$e^{\sqrt{d}\log d} \leq n \leq e^{d\log d/2},$$

$$p = \left\lceil \frac{4\log(m_0)}{\log(d/m_0)} \right\rceil$$

and

$$\sqrt{d} - 1 \leq m_0 \leq d/2.$$

By our choice of p,

$$\left(\frac{m_0}{d}\right)^p \leq \left(\frac{m_0}{d}\right)^{\frac{4\log(m_0)}{\log(d/m_0)}} = \frac{1}{m_0^4}$$

which concludes the proof. □

As can be seen from the proof of Theorem 1.2, the error term in (16) may be improved to $O(1/d^5)$ or so if we take $m = m_0 12$ rather than $m = m_0 - 4$. In any case, the error term is typically negligible compared to $\Psi^2_\mu(m)$.

3.2 Least square estimator

We now move to the analysis of the least squares estimator $\hat{f}_{LS}$. Given a choice of a polynomial degree m, this estimator is defined as the polynomial of degree less than m that minimizes the empirical l^2 risk:

$$\hat{f}_{LS} = \operatorname*{argmin}_{P \in \mathcal{P}_{d,m}} \sum_{i=1}^n \left(Y_i - P(X_i)\right)^2. \tag{54}$$

The goal of this section is to prove the following bound on its prediction error:

Theorem 3.2. *Assume that*
$$\sigma^2 \leq d$$
and that $d \geq d_0$, for some universal constant $d_0 \geq 3$. Then for any $n, m \geq 1$ such that the right-hand side is smaller than 1, it holds that

$$\mathbb{E}\|f - \hat{f}_{LS}\|_{L^2(\mu)}^2 \leq \Psi_\mu^2(m) + \frac{(C\log(n))^{2m}D\log(D)}{n} + \frac{8\sigma^2 D}{n}, \tag{55}$$

for some absolute constant $C \geq 1$.

Before embarking on the proof of Theorem 3.2, we remark that the assumptions implies in particular that

$$\log(n)^m D \leq (C\log(n))^{2m}D \leq n. \tag{56}$$

Thus, if $n \geq 3$, we get

$$m \leq \log n. \tag{57}$$

Furthermore,

$$D = \binom{d+m}{m} \geq \left(\frac{d}{m}\right)^m. \tag{58}$$

Plugging (58) and (57) into (56) we get

$$d^m \leq n,$$

that is

$$m \leq \frac{\log n}{\log d}. \tag{59}$$

which we assume in the rest of this section. Note that the quantity of interest,

$$f - \hat{f}_{LS},$$

is unchanged if we subtract a constant from f. Thus, for convenience, we assume from now on in this section that

$$\int f \, d\mu = 0. \tag{60}$$

We define

$$A = (p_k(X_i))_{k,i} = \begin{pmatrix} p_0(X_1) & \cdots & p_{D-1}(X_1) \\ \vdots & & \vdots \\ p_0(X_n) & \cdots & p_{D-1}(X_n) \end{pmatrix} \in \mathbb{R}^{n \times D},$$

and

$$b = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

We adopt the following useful notation: for a polynomial P of degree at most m, we write

$$\boldsymbol{P} \in \mathbb{R}^D$$

for the vector of its coordinates in the basis $(p_k)_{0 \leq k \leq D-1}$. A straightforward computation shows that

$$\hat{\boldsymbol{f}}_{\boldsymbol{LS}} = (A^T A)^{-1} A^T b = \left(\frac{1}{n}A^T A\right)^{-1} \frac{1}{n} A^T b. \tag{61}$$

The vector $\frac{1}{n}A^T b$ is merely the vector of empirical scalar products, which, as in the previous section, we denote by $\hat{\boldsymbol{f}}^*$. For all $0 \le k \le D - 1$,

$$\hat{\boldsymbol{f}}_{\boldsymbol{k}}^* = \left(\frac{1}{n}A^T b\right)_k = \frac{1}{n}\sum_{i=1}^n Y_i p_k(X_i). \tag{62}$$

This is indeed the same definition as in (48), since we assumed that

$$a = \int f \, d\mu = 0.$$

From the analysis carried out in Section 3.1, we know that

$$\mathbb{E}\|\boldsymbol{P_m f} - \hat{\boldsymbol{f}}^*\|_2^2 \le \frac{(Cm^2 + 4\sigma^2)D}{n} \tag{63}$$

where $P_m f$ is the projection of f onto $\mathcal{P}_{d,m}$ in $L^2(\mu)$. From now on, we assume that n is large enough so that the right hand side in (55) is smaller than 1. In particular, we also get

$$\mathbb{E}\|\boldsymbol{P_m f} - \hat{f}^*\|_{L^2(\mu)}^2 \lesssim 1.$$

Since $\|f\|_{L^2(\mu)} \le \Psi_\mu(0) = 1$,

$$\|\hat{f}^*\|_{L^2(\mu)} \le \|\boldsymbol{P_m f} - \hat{f}^*\|_{L^2(\mu)} + \|\boldsymbol{P_m f}\|_{L^2(\mu)} \le C + \|f\|_{L^2(\mu)} \le \tilde{C}. \tag{64}$$

We denote

$$C_n = \frac{1}{n}A^T A.$$

The main technical step in this section is a moment bound on the deviation of C_n^{-1} from the identity matrix, measured in operator norm.

Lemma 3.3. *Assume that $n \ge D$, then for all $1 \le p \le \log D$,*

$$\left(\mathbb{E}\|C_n^{-1} - I_d\|_{op}^p\right)^{2/p} \le \frac{(C \log n)^{2m} D \log D}{n}$$

where $C > 0$ is a universal constant.

Before proving this lemma, let us explain how it implies Theorem 3.2. As a warm-up, we first prove a weaker statement:

$$\mathbb{E}\|f - \hat{f}_{LS}\|_2 \le \Psi_\mu(m) + \sqrt{\frac{(C \log n)^{2m} D \log D}{n}} + \frac{2\sigma\sqrt{D}}{\sqrt{n}}. \tag{65}$$

3.2.1   Proof of (65)

Since $\hat{\boldsymbol{f}}^*$ is given by (62), we may write

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{P_m f} - \hat{\boldsymbol{f}}_{\boldsymbol{LS}}\|_2 &\le \mathbb{E}\|\boldsymbol{P_m f} - \hat{\boldsymbol{f}}^*\|_2 + \mathbb{E}\|\hat{\boldsymbol{f}}_{\boldsymbol{LS}} - \hat{\boldsymbol{f}}^*\|_2 \\
&= \mathbb{E}\|\boldsymbol{P_m f} - \hat{\boldsymbol{f}}^*\|_2 + \mathbb{E}\|(C_n^{-1} - I_d)\,\hat{\boldsymbol{f}}^*\|_2 \\
&\le \mathbb{E}\|\boldsymbol{P_m f} - \hat{\boldsymbol{f}}^*\|_2 + \mathbb{E}\|C_n^{-1} - I_d\|_{op}\,\|\hat{\boldsymbol{f}}^*\|_2.
\end{aligned}$$

Now, using the Cauchy–Schwarz inequality and (64), we bound the last term by

$$\mathbb{E}\|C_n^{-1} - I_d\|_{op}\,\|\hat{\boldsymbol{f}}^*\|_2 \lesssim \left(\mathbb{E}\|C_n^{-1} - I_d\|_{op}^2\right)^{1/2}.$$

23

From Lemma 3.3 with $p = 2$, we know that

$$\mathbb{E}\|C_n^{-1} - I_d\|_{op}^2 \lesssim \frac{(C \log n)^{2m} D \log D}{n}. \tag{66}$$

Putting everything together and using (63), we get

$$\mathbb{E}\|f - \hat{f}_{LS}\|_{L^2(\mu)} \leq \|f - P_m f\|_{L^2(\mu)} + \mathbb{E}\|P_m f - \hat{f}_{LS}\|_{L^2(\mu)}$$
$$\leq \Psi_\mu(m) + \sqrt{\frac{(Cm^2 + 4\sigma^2)D}{n}} + \sqrt{\frac{(C \log n)^{2m} D \log D}{n}}$$
$$\leq \Psi_\mu(m) + \sqrt{\frac{(C \log n)^{2m} D \log D}{n}} + \frac{2\sigma\sqrt{D}}{\sqrt{n}},$$

where the constant C may change from line to line, and we used $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$.

3.2.2 Proof of Theorem 3.2

The proof of (55) follows the same strategy, with one additional computation. As before, we write

$$\mathbb{E}\|f - \hat{f}_{LS}\|_{L^2(\mu)}^2 = \mathbb{E}\|f - P_m f\|_{L^2(\mu)}^2 + \mathbb{E}\|P_m f - \hat{f}_{LS}\|_{L^2(\mu)}^2$$
$$\leq \Psi_\mu^2(m) + 2\,\mathbb{E}\|\boldsymbol{P_m f} - \hat{\boldsymbol{f}}^*\|_2^2 + 2\,\mathbb{E}\|\hat{\boldsymbol{f}}_{LS} - \hat{\boldsymbol{f}}^*\|_2^2$$
$$\leq \Psi_\mu^2(m) + \frac{(Cm^2 + 8\sigma^2)D}{n} + 2\,\mathbb{E}\|C_n^{-1} - I_d\|_{op}^2 \|\hat{\boldsymbol{f}}^*\|_2^2.$$

Using Hölder's inequality with $p = \frac{1}{2} \log D$ and $q = p^*$, and using Lemma 3.3, we bound

$$\mathbb{E}\|C_n^{-1} - I_d\|_{op}^2 \|\hat{\boldsymbol{f}}^*\|_2^2 \leq \left(\mathbb{E}\|C_n^{-1} - I_d\|_{op}^{2p}\right)^{1/p} \left(\mathbb{E}\|\hat{\boldsymbol{f}}^*\|_2^{2q}\right)^{1/q}$$
$$\leq \frac{(C \log n)^{2m} D \log D}{n} \left(\mathbb{E}\|\hat{\boldsymbol{f}}^*\|_2^{2q}\right)^{1/q}.$$

It remains to prove that

$$\left(\mathbb{E}\|\hat{\boldsymbol{f}}^*\|_2^{2q}\right)^{1/q} \lesssim 1. \tag{67}$$

We use a simple interpolation argument. Recall that, by (64),

$$\mathbb{E}\|\hat{\boldsymbol{f}}^*\|_2^2 \lesssim 1. \tag{68}$$

Recall that $\sigma^2 \leq d \leq D$. We claim the following crude bound on the fourth moment:

$$\mathbb{E}\|\hat{\boldsymbol{f}}^*\|_2^4 \lesssim D^3. \tag{69}$$

Indeed,

$$\mathbb{E}\|\hat{\boldsymbol{f}}^*\|_2^4 = \mathbb{E}\left(\sum_{k=0}^{D-1} (\hat{f}_k^*)^2\right)^2 \leq D \sum_{k=0}^{D-1} \mathbb{E}(\hat{f}_k^*)^4.$$

For any $0 \leq k \leq D - 1$,

$$\mathbb{E}((\hat{f}_k^*)^4) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n Y_i\, p_k(X_i)\right)^4 \leq \mathbb{E}(Y_1^4\, p_k(X_1)^4)$$
$$\leq (\mathbb{E}Y_1^8)^{1/2} (\mathbb{E}|p_k(X_1)|^8)^{1/2} \lesssim \sigma^4 C^m \lesssim D^2,$$

24

where we used Propositions 2.1 and 2.3 and the growth assumption on σ. Now, for any nonnegative random variable X and any $q \in [1, 2]$, by interpolation,

$$\left(\mathbb{E}X^q\right)^{1/q} \leq \left(\mathbb{E}X\right)^{2/q-1}\left(\mathbb{E}X^2\right)^{1-1/q}.$$

Applying this to $X = \|\hat{f}^*\|_2^2$ and using (68) and (69), we get

$$\left(\mathbb{E}\|\hat{f}^*\|_2^{2q}\right)^{1/q} \lesssim (D^3)^{1-1/q} = D^{3/p} = D^{6/\log D} \lesssim 1,$$

which proves (67). In order to complete the proof of Theorem 3.2, it remains to prove Lemma 3.3.

3.2.3 Bounding C_n^{-1}

The proof of Lemma 3.3 consists of two steps. First, we prove the same bound but for C_n rather than for its inverse.

Lemma 3.4. *Assume that $n \geq D$, then for all $1 \leq p \leq \log D$,*

$$\left(\mathbb{E}\|C_n - I_d\|_{op}^p\right)^{2/p} \leq \frac{(C\log n)^{2m}D\log D}{n}.$$

where $C > 0$ is a universal constant.

In order to prove Lemma 3.4, we first unpack the definition of C_n. For $1 \leq i \leq n$, define i.i.d random vectors

$$Z_i = \begin{pmatrix} p_1(X_i) \\ \vdots \\ p_{D-1}(X_i) \end{pmatrix} \in \mathbb{R}^{D-1} \qquad (i = 1, \ldots, n). \tag{70}$$

Notice that we did not include the term $p_0(X_i) = 1$ corresponding to the constant polynomial. We write Z for a random vector with the same law as Z_1. Then Z is isotropic:

$$\mathbb{E}Z = 0, \qquad \mathbb{E}ZZ^\top = I_d.$$

We denote the empirical covariance of Z by

$$\mathrm{Cov}_n = \frac{1}{n}\sum_{i=1}^n Z_iZ_i^\top. \tag{71}$$

We also write $\tilde{Z}_i$ for the full vector

$$\tilde{Z}_i = \begin{pmatrix} 1 \\ Z_i \end{pmatrix} \in \mathbb{R}^D. \tag{72}$$

We may then express C_n as

$$C_n = \frac{1}{n}AA^\top = \frac{1}{n}\sum_{i=1}^n \tilde{Z}_i\tilde{Z}_i^\top$$

$$= \begin{pmatrix} 1 & \frac{1}{n}\sum_{i=1}^n Z_i^\top \\ \frac{1}{n}\sum_{i=1}^n Z_i & \mathrm{Cov}_n \end{pmatrix}. \tag{73}$$

From (73) and (72), we easily deduce the following lemma.

Lemma 3.5. *Let C_n and Cov_n be defined by (73) and (71), respectively. Then,*

$$\|C_n - I_d\|_{op} \leq \left\|\frac{1}{n}\sum_{i=1}^n Z_i\right\| + \|\mathrm{Cov}_n - I_d\|_{op}. \tag{74}$$

25

In what follows, we bound the p-th moment of the operator norm of $\mathrm{Cov}_n - I_d$. We use Rudelson's lemma [Rud99], relying on the non-commutative Khintchine inequality of Lust-Picard and Pisier (see [Pis03, Theorem 9.8.2]). Inequality (3.4) in [Rud99] reads as follows:

Lemma 3.6. *Let $x_1, \ldots, x_n$ be vectors in $\mathbb{R}^D$, and let $\epsilon_1, \ldots, \epsilon_n$ be i.i.d. symmetric Bernoulli variables. Then for any $p \leq \log D$,*

$$\left(\mathbb{E}\|\sum_i \epsilon_i x_i \otimes x_i\|_{op}^p\right)^{2/p} \leq C \log D \, \max_i \|x_i\|^2 \, \|\sum_i x_i \otimes x_i\|_{op}.$$

As in Rudelson's paper, the lemma is used to bound the deviation of the empirical covariance from its expectation.

Corollary 3.7. *Let Cov_n be defined by (71). Whenever the right-hand side is smaller than 1,*

$$\left(\mathbb{E}\|\mathrm{Cov}_n - I_d\|^p\right)^{2/p} \leq \frac{C \log D}{n} \left(\mathbb{E} \max_i |Z_i|^{2p}\right)^{1/p}.$$

We need a standard symmetrization lemma.

Lemma 3.8. *Let $(X_i)_{i \in I}$ be a finite sequence of independent random vectors in some Banach space, and let ε_i be independent symmetric Bernoulli random variables. Then, for any $p \geq 1$,*

$$\mathbb{E}\|\sum_{i \in I} X_i - \mathbb{E}X_i\|^p \leq 2^p \, \mathbb{E}\|\sum_{i \in I} \varepsilon_i X_i\|^p.$$

Proof. We set
$$\tilde{X}_i = X_i - X_i',$$

where X_i' is an independent copy of X_i. By Jensen's inequality,

$$\begin{aligned}
\mathbb{E}\|\sum_{i \in I} X_i - \mathbb{E}X_i\|^p &\leq \mathbb{E}\|\sum_{i \in I} \tilde{X}_i\|^p \\
&= \mathbb{E}\|\sum_{i \in I} \varepsilon_i \tilde{X}_i\|^p \\
&\leq 2^{p-1}\mathbb{E}\left(\|\sum_{i \in I} \varepsilon_i X_i\|^p + \|\sum_{i \in I} \varepsilon_i X_i'\|^p\right) \\
&= 2^p \, \mathbb{E}\|\sum_{i \in I} \varepsilon_i X_i\|^p.
\end{aligned}$$

$\square$

We can now prove Corollary 3.7.

Proof of Corollary 3.7. We set
$$S_p = \mathbb{E}\|\mathrm{Cov}_n - I_d\|_{op}^p,$$

the quantity of interest. The first step is to use the symmetrization lemma:

$$\begin{aligned}
S_p &= \mathbb{E}\|\frac{1}{n}\sum_{i=1}^n (Z_i \otimes Z_i - \mathbb{E}Z_i \otimes Z_i)\|_{op}^p \\
&\leq \frac{2^p}{n^p}\mathbb{E}\|\sum_{i=1}^n \varepsilon_i Z_i \otimes Z_i\|_{op}^p.
\end{aligned}$$

We then apply Rudelson's lemma, conditionally on the Z_i's and then take expectation over the Z_i's, to obtain

$$S_p \le \frac{C^p}{n^p} \log(D)^{p/2} \, \mathbb{E}\left(\max_i \|Z_i\|^p \, \|\sum_i Z_i \otimes Z_i\|_{op}^{p/2} \right)$$

$$\le \frac{C^p}{n^p} \log(D)^{p/2} \left(\mathbb{E} \max_i \|Z_i\|^{2p} \right)^{1/2} \left(\mathbb{E}\|\sum_i Z_i \otimes Z_i\|_{op}^p \right)^{1/2}, \tag{75}$$

where we used Cauchy–Schwarz. Now, observe that

$$\mathbb{E}\|\sum_i Z_i \otimes Z_i\|_{op}^p = n^p \, \mathbb{E}\|I_d + \frac{1}{n}\sum_i (Z_i \otimes Z_i - I_d)\|_{op}^p$$

$$\le 2^{p-1} n^p \left(1 + \mathbb{E}\|\frac{1}{n}\sum_i (Z_i \otimes Z_i - I_d)\|_{op}^p \right)$$

$$= 2^{p-1} n^p (1 + S_p).$$

Plugging this back into (75), we find that

$$S_p \le \lambda (1 + S_p)^{1/2}, \tag{76}$$

where

$$\lambda = \frac{(C \log D)^{p/2}}{n^{p/2}} \left(\mathbb{E} \max_i |Z_i|^{2p} \right)^{1/2}.$$

Distinguishing the cases $S_p \le 1$ and $S_p \ge 1$, one obtains

$$S_p \le 2\max(\lambda, \lambda^2).$$

Thus, in particular, if $\lambda \le 1$, we conclude that

$$\left(\mathbb{E}\|\mathrm{Cov}_n - I_d\|_{op}^p \right)^{2/p} \le \lambda^{2/p} = \frac{C \log D}{n} \left(\mathbb{E} \max_i |Z_i|^{2p} \right)^{1/p},$$

which is the desired bound. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

We now need an estimate on

$$\left(\mathbb{E} \max_i |Z_i|^{2p} \right)^{1/p}.$$

The ℓ_∞ norm on $\mathbb{R}^n$ is equivalent to the ℓ_q norm for $q = 2 \log n$, up to a universal constant. For this choice of q, notice that $2p/q \le 1$. By Jensen's inequality,

$$\left(\mathbb{E} \max_i \|Z_i\|^{2p} \right)^{1/p} \lesssim \left(\mathbb{E}\left(\sum_{i=1}^n |Z_i|^q \right)^{2p/q} \right)^{1/p}$$

$$\lesssim \left(n \, \mathbb{E}|Z_1|^q \right)^{2/q}$$

$$\lesssim \left(\mathbb{E}|Z_1|^q \right)^{2/q}.$$

Finally, the random variable $Q = |Z_1|^2$ is a degree $2m$ polynomial in log-concave variables with

$$\mathbb{E}|Q| = \mathbb{E}Q = \mathbb{E}|Z_1|^2 = D - 1.$$

By Proposition 2.3, we obtain

$$
\begin{aligned}
\left(\mathbb{E}|Z_1|^q\right)^{2/q} &= \left(\mathbb{E}Q^{q/2}\right)^{2/q} \\
&\leq (Cq/2)^{2m}\, D \\
&= D(C\log n)^{2m}.
\end{aligned}
$$

At this point, we have established the bound (whenever the right-hand side is smaller than 1):

$$
\left(\mathbb{E}\|\mathrm{Cov}_n - I_d\|_{op}^p\right)^{2/p} \leq \frac{(C\log n)^{2m}D\log D}{n}. \tag{77}
$$

In view of Lemma 3.5, we have, for any $p \geq 1$,

$$
\begin{aligned}
\left(\mathbb{E}\|C_n - I_d\|_{op}^p\right)^{2/p} &\leq 2\left(\mathbb{E}\|\mathrm{Cov}_n - I_d\|_{op}^p\right)^{2/p} + 2\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n Z_i\right|_2^p \\
&\leq \frac{(C\log n)^{2m}D\log D}{n} + 2\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n Z_i\right|_2^p.
\end{aligned} \tag{78}
$$

Thus, we need to upper-bound

$$
\mathbb{E}\|\frac{1}{n}\sum_{i=1}^n Z_i\|_2^p.
$$

First, for $p = 2$ we have

$$
\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n Z_i\right|^2 = \frac{D-1}{n} \leq \frac{D}{n}.
$$

For general p, consider the random variable

$$
\tilde{Q} = \left|\frac{1}{n}\sum_{i=1}^n Z_i\right|_2^2.
$$

It is a polynomial of degree $2m$ in the log-concave variables X_i. Furthermore, from the case $p = 2$, we know that

$$
\mathbb{E}|\tilde{Q}| = \mathbb{E}\tilde{Q} \leq \frac{D}{n}.
$$

Using again the moment inequality for polynomials (Proposition 2.3), we find that

$$
\left(\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n Z_i\right|_2^p\right)^{2/p} = \left(\mathbb{E}\tilde{Q}^{p/2}\right)^{2/p} \leq (Cp)^{2m}\mathbb{E}|\tilde{Q}| \leq \frac{(C\log n)^{2m}D}{n},
$$

where we used that $D \leq n$. Plugging this inequality into (78) finally proves Lemma 3.4:

$$
\left(\mathbb{E}\|C_n - I_d\|_{op}^p\right)^{2/p} \leq \frac{(C\log n)^{2m}D\log D}{n}.
$$

$\square$

It remains to pass from an inequality on C_n to a corresponding inequality for its inverse. We shall thus need an integrable bound on the probability that the smallest eigenvalue of C_n is small.

Recall that the covariance matrix is given by

$$
C_n = \frac{1}{n}\sum_{i=1}^n \tilde{Z}_i \otimes \tilde{Z}_i,
$$

28

where $\tilde{Z}_i = (p_k(X_i))_{0 \leq k \leq D-1}$. For any $\theta \in \mathbb{S}^{D-1}$, $\tilde{Z}_i \cdot \theta$ is a polynomial of degree at most m with

$$\mathbb{E}|\tilde{Z}_i \cdot \theta|^2 = 1.$$

The Carbery-Wright Theorem (2.5) implies the following small-ball property.

Lemma 3.9. *For any $\theta \in \mathbb{S}^{D-1}$ and any $t \geq 0$,*

$$\mathbb{P}(|\tilde{Z} \cdot \theta| \leq t) \leq Cm\, t^{1/m}.$$

In the sequel, we work in the setting of Theorem 3.2. In particular, we may assume that $n \geq C_0^m D$ for some sufficiently large constant C_0 and $m \geq 1$. We control the tails of $\lambda_{\min}(C_n)$ in two regimes:

Lemma 3.10. *Assume as we may that $n \geq C_0^m D$ for some sufficiently large C_0. Then there exist universal constants c_0, c_1, c_2 such that*

$$\mathbb{P}(\lambda_{\min}(C_n) \leq e^{-c_0 m}) \leq \exp\left(-\frac{n}{e^{c_1 m}}\right).$$

Furthermore, for $t \leq 1/n^2$,

$$\mathbb{P}(\lambda_{\min}(C_n) \leq t) \leq t^{n/16m}.$$

Proof. We start by proving the first statement. Notice that

$$\lambda_{\min}(C_n) = \inf_{\theta \in S^{D-2}} \frac{1}{n} \sum_{i=1}^{n} (\tilde{Z}_i \cdot \theta)^2.$$

Now fix some unit vector θ, and write $V = (\tilde{Z} \cdot \theta)^2$. Then V is a non-negative random variable with $\mathbb{E}V = 1$ and $\mathbb{E}V^2 \leq e^{cm}$ for some constant c, by Proposition 2.3. The Paley–Zygmund inequality implies that

$$\mathbb{P}(V < 1/2) \leq 1 - e^{-\tilde{c}m}$$

for some constant $\tilde{c} > 0$. This in turn implies that

$$\mathbb{E}e^{-mV} \leq 1 - e^{-cm} \tag{79}$$

for some constant $c > 0$. We now make use of the Laplace method. Fix $\theta \in \mathbb{S}^{D-1}$ and write

$$S_n = \sum_{i=1}^{n} (\tilde{Z}_i \cdot \theta)^2 = \sum_{i=1}^{n} V_i.$$

Let $c_0 = c + \log 2$, and let $t_1 = e^{-c_0 m}$. By Markov's inequality,

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n} S_n < t_1\right) &= \mathbb{P}\left(e^{-mS_n} > e^{-nmt_1}\right) \\
&\leq \left(\mathbb{E}(e^{-mV})e^{mt_1}\right)^n \\
&\leq \left((1 - e^{-cm})(1 + 2e^{-c_0 m})\right)^n \\
&\leq (1 - e^{-2cm})^n \\
&\leq \exp\left(-\frac{n}{e^{2cm}}\right),
\end{aligned}$$

where we used that $mt_1 \leq me^{-m\log 2} \leq 1$ and that $e^x \leq 1 + 2x$ for $x \leq 1$ and that $2e^{-c_0 m} \leq e^{-cm}$.

Now taking a union bound over a $t_1/2$-net $\mathcal{N}$ of the sphere $\mathbb{S}^{D-2}$ of cardinality

$$|\mathcal{N}| \leq \left(1 + \frac{4}{t_1}\right)^D$$

concludes the proof of the first statement, since $n \geq C_0^m D$ for a sufficiently large chosen C_0.

We move to the second statement. Again, we fix some vector on the sphere and work with the same notations as before. For any $t \leq \frac{1}{n^2}$,

$$
\begin{aligned}
\mathbb{P}\left(\tfrac{1}{n} S_n \leq 2t\right) &\leq \mathbb{P}(S_n \leq 2\sqrt{t})^n \\
&\leq \mathbb{P}(V \leq 2\sqrt{t})^n \\
&= \mathbb{P}(|Z \cdot \theta| \leq t^{1/4})^n \\
&\leq Cm\, t^{n/4m} \\
&\leq t^{n/8m},
\end{aligned}
$$

where we used Carbery–Wright (Theorem 2.5) on line 4 and assumed a large enough choice of C_0. Taking a union bound over a t-net of the sphere, of cardinality less than $(1 + 2/t)^D$, concludes the proof, again for C_0 large enough.

We are now in a position to prove Lemma 3.3. We use the simple observations that, for any positive matrix M,

$$\|M^{-1} - I_d\|_{op} \leq \frac{1}{\lambda_{\min}(M)}\|M - I_d\|_{op},$$

and that

$$\|M^{-1} - I_d\|_{op} \leq \max\left(1, \frac{1}{\lambda_{\min}(M)}\right).$$

We abbreviate $\lambda_{\min} = \lambda_{\min}(C_n)$. Recall that

$$m \leq \frac{\log n}{\log d}.$$

Thus, given c_0 and c_1 the constants from Lemma 3.10, if d is large enough we have

$$e^{-c_0 m} \geq \frac{1}{n^2}, \qquad e^{c_1 m} \leq \sqrt{n}.$$

We partition the probability space into three events:

$$\mathcal{A} = \{\lambda_{\min} \geq e^{-c_0 m}\},$$
$$\mathcal{B} = \left\{\tfrac{1}{n^2} \leq \lambda_{\min} \leq e^{-c_0 m}\right\},$$
$$\mathcal{C} = \{\lambda_{\min} \leq \tfrac{1}{n^2}\}.$$

Using the previous observations and Lemma 3.10,

$$
\begin{aligned}
\mathbb{E}\|C_n^{-1} - I_d\|_{op}^p &= \mathbb{E}\big[\|C_n^{-1} - I_d\|_{op}^p \mathbb{1}_{\mathcal{A}}\big] + \mathbb{E}\big[\|C_n^{-1} - I_d\|_{op}^p \mathbb{1}_{\mathcal{B}}\big] + \mathbb{E}\big[\|C_n^{-1} - I_d\|_{op}^p \mathbb{1}_{\mathcal{C}}\big] \\
&\leq e^{pc_0 m}\mathbb{E}\big[\|C_n - I_d\|_{op}^p \mathbb{1}_{\mathcal{A}}\big] + \mathbb{E}\big[\lambda_{\min}^{-p} \mathbb{1}_{\mathcal{B}}\big] + \mathbb{E}\big[\lambda_{\min}^{-p} \mathbb{1}_{\mathcal{C}}\big] \\
&\leq e^{pc_0 m}\mathbb{E}\|C_n - I_d\|_{op}^p + n^{2p}\exp\left(-\frac{n}{e^{c_1 m}}\right) + \int_{n^{2p}}^{+\infty} \mathbb{P}\left(\frac{1}{\lambda_{\min}^p} \geq u\right) du \\
&\leq e^{pc_0 m}\mathbb{E}\|C_n - I_d\|_{op}^p + n^{2p}e^{-\sqrt{n}} + \int_0^{1/n^2} \mathbb{P}(\lambda_{\min} \leq t)\frac{p}{t^{p+1}} dt \\
&\leq e^{pc_0 m}\mathbb{E}\|C_n - I_d\|_{op}^p + O(e^{-n^{1/4}}) + p\int_0^{1/n^2} t^{n/16m - p - 1}\, dt \\
&\leq e^{pc_0 m}\mathbb{E}\|C_n - I_d\|_{op}^p + O(e^{-n^{1/4}}).
\end{aligned}
$$

We conclude that
$$\left(\mathbb{E}\|C_n^{-1} - I_d\|_{op}^p\right)^{2/p} \leq \frac{(C\log n)^{2m}D\log D}{n}$$
for some constant $C > 0$, which is what we wanted. $\qquad\square$

3.2.4 Proof of Theorem 1.4

We explain how to deduce Theorem 1.4 from Theorem 3.2. As in 3.1.1, we set
$$m_0 = \left\lfloor \frac{\log n}{\log d} \right\rfloor,$$
and for any integer $1 \leq p \leq m_0$,
$$m_p = m_0 - p.$$
We again have
$$\frac{\binom{d+m_p}{m_p}}{n} \leq \left(\frac{m_0}{d}\right)^p.$$
Plugging this into Theorem 3.2, we obtain, for the choice $m = m_p$,

$$
\begin{aligned}
\mathbb{E}\|f - f_{LS}\|_{L^2(\mu)}^2 &\leq \Psi_\mu^2(m_p) + \frac{(C\log n)^{2m_p}D\log D}{n} + \frac{8\sigma^2 D}{n} \\
&\leq \Psi_\mu^2(m_p) + (C\log n)^{2m_0+1-p}\left(\frac{m_0}{d}\right)^p + 8d\left(\frac{m_0}{d}\right)^p \\
&\leq \Psi_\mu^2(m_p) + (C\log n)^{2m_0+1}\left(\frac{1}{d\log d}\right)^p + 8d\left(\frac{m_0}{d}\right)^p.
\end{aligned}
$$

First regime. We set $p = 4$ and assume that
$$n \leq \exp\left(\frac{c\log^2 d}{\log\log d}\right)$$
for some constant $c < 1$ to be determined. As a consequence, we have
$$\log n \cdot \log\log n \leq 2c\log^2 d,$$
where we assume $d \geq 16$. We upper bound

$$
\begin{aligned}
(C\log n)^{2m+1} &\leq (C\log n)^{\frac{4c\log d}{\log\log n}+1} \\
&\leq \exp\left(\left(\frac{4c\log d}{\log\log n}+1\right)\cdot\log(C\log n)\right) \\
&\leq \exp\left(\left(\frac{4c\log d}{\log\log n}+1\right)(\log\log n + C')\right) \\
&\leq \exp\left(4c\log d + \log\log n + \frac{4c\log d}{\log\log n}C' + C'\right) \\
&\leq \exp(4c(1+C')\log d + 2\log\log d + C') \\
&\leq \exp(C' + 2\log d) \\
&\lesssim d^2,
\end{aligned}
$$

where we chose
$$c = \frac{1}{4(1 + C')}.$$

On the other hand, clearly
$$8d \left(\frac{m_0}{d}\right)^p \leq 8d \left(\frac{\log d}{d}\right)^4 = o(1/d).$$

Second regime. We want to ensure that
$$(C \log n)^{2m_0+1} \leq (d \log d)^{p-1}.$$

We assume that
$$\alpha := \frac{\log(C \log n)}{\log d} < 1/2$$

It is enough that
$$p \geq 1 + \frac{2m_0 \log(C \log n)}{\log d} + \frac{\log(C \log n)}{\log d}.$$

Thus, using that $\alpha < 1$, it is enough that
$$p \geq 2 + \frac{2m_0 \log(C \log n)}{\log d}.$$

As announced, we choose
$$p = 4 + \left\lfloor \frac{2m_0 \log(C \log n)}{\log d} \right\rfloor$$
$$= 4 + \lfloor 2\alpha m_0 \rfloor.$$

For that choice of p, since
$$m_0 = \left\lfloor \frac{\log n}{\log d} \right\rfloor \lesssim d^{1/2},$$

we again have
$$\sigma^2 \frac{D}{n} \leq d \left(\frac{m_0}{d}\right)^4 = O(1/d).$$

4 The metric entropy of Lipschitz functions

In the previous sections, we used low-degree multivariate polynomials to approximate and learn Lipschitz functions in high dimensions. In the Gaussian setting, when $\mu = \gamma$, for a given $\varepsilon > 0$ we use that any 1-Lipschitz function f can be approximated with error at most ε by a polynomial of degree at most m, where $m \simeq \frac{1}{\varepsilon}$. Heuristically, this approach makes sense if, at scale ε, the "size" of the space of polynomials of degree at most m is not much larger than that of the space of Lipschitz functions. One standard way to measure size is through metric entropy. For a metric space $(\mathcal{X}, d)$ we define its metric entropy as
$$H_{(\mathcal{X},d)}(\varepsilon) = \log \mathcal{N}_{(\mathcal{X},d)}(\varepsilon), \tag{80}$$
for all $\varepsilon > 0$, where $\mathcal{N}_{(\mathcal{X},d)}(\varepsilon)$ is the largest cardinality of an ε-separated set in $(\mathcal{X}, d)$. We adopt the (slightly unusual) convention of using packing numbers instead of covering numbers for our definition of metric entropy, as it will be more convenient for us.

We provide estimates for the metric entropy of Lipschitz functions equipped with the distance
$$d(f, g) = \|f - g\|_{L^2(\mu)},$$

where μ is an isotropic product log-concave measure on $\mathbb{R}^d$. We denote by

$$H_L^\mu(\varepsilon) = H_{(B_{Lip}^\mu, d)}(\varepsilon),$$

where B_{Lip}^μ is the unit ball of 1-Lipschitz functions, i.e., those f such that

$$\int f^2 \, d\mu \leq 1.$$

Theorem 4.1. *Let μ be an isotropic product log-concave measure on $\mathbb{R}^d$. Let $\varepsilon > 0$ satisfy*

$$d^{-1/4} < \varepsilon < 1.$$

Then

$$d^{c/\varepsilon^2} \lesssim H_L^\mu(\varepsilon),$$

where $c > 0$ is a universal constant.

In the case where $\mu = \gamma$, the standard Gaussian measure on $\mathbb{R}^d$, we get a two-sided estimate:

Corollary 4.2. *There exists a constant $c > 0$ such that, for any ε with*

$$d^{-1/4} < \varepsilon < 1,$$

we have

$$\binom{d}{\lfloor c/\varepsilon \rfloor^2} \lesssim H_L^\gamma(\varepsilon) \lesssim \binom{d}{\lceil 4/\varepsilon \rceil^2}, \tag{81}$$

where $c > 0$ is a universal constant.

Remark 4.3. Corollary 4.2 extends immediately to products of isotropic log-concave measures that are Lipschitz images of the Gaussian. Indeed, if $\mu = T \# \gamma$ for some K-Lipschitz map T, then

$$d_\mu(f, g) = d_\gamma(f \circ T, g \circ T)$$

and thus

$$H_L^\mu(\varepsilon) \leq H_L^\gamma(\varepsilon/K).$$

This includes, for example, the uniform measure on the hypercube, or products of strongly log-concave densities.

From Theorem 4.1, we will deduce a slightly weaker lower bound for the general case.

Theorem 4.4. *Let μ be an isotropic log-concave probability measure on $\mathbb{R}^d$. Let $\varepsilon > 0$ satisfy*

$$d^{-\eta} < \varepsilon < 1.$$

Then

$$d^{c/\varepsilon^2} \lesssim H_L^\mu(\varepsilon),$$

where $\eta < 1/4$ and $c > 0$ are universal constants.

We begin by proving the upper bound in Corollary 4.2, which essentially follows from polynomial approximation. Without loss of generality, we may assume that d is large enough. Let $\varepsilon \in (0, 1)$ and let $(f_1, \ldots, f_N)$ be an ε-separated subset of B_{Lip}^γ. Since $\Psi_\gamma(m) \leq 1/(m+1)$, there exist polynomials $P_1, \ldots, P_N$ such that

$$\|f_i - P_i\|_{L^2(\gamma)} \leq \tfrac{\varepsilon}{3}, \qquad \deg(P_i) \leq m := \left\lceil \tfrac{3}{\varepsilon} \right\rceil^2.$$

Thus, by the triangle inequality, the set $(P_1, \ldots, P_N)$ is $\varepsilon/3$-separated; indeed, for $i \neq j$,

$$\|P_i - P_j\|_{L^2(\gamma)} \geq \tfrac{\varepsilon}{3}.$$

In fact, for any i, the polynomial P_i is the truncated Hermite expansion of the 1-Lipschitz function f_i:

$$P_i = \sum_{|\alpha| \leq m} \langle f_i, H_\alpha \rangle \, H_\alpha.$$

In particular,

$$\|P_i\|_{L^2(\gamma)} \leq \|f_i\|_{L^2(\gamma)} \leq 1.$$

Hence P_i lies in the unit ball of $\mathcal{P}_{d,m}$, equipped with the norm $\|\cdot\|_{L^2(\gamma)}$. As before, we set

$$D = \binom{d+m}{m}$$

for the dimension of that space. We thus have the standard packing bound

$$N \leq \left(1 + \tfrac{6}{\varepsilon}\right)^D \leq \left(\tfrac{7}{\varepsilon}\right)^D.$$

Let $m_2 := \left\lceil \tfrac{4}{\varepsilon} \right\rceil^2$ and $D_2 := \binom{d+m_2}{m_2}$. For d large enough,

$$D \log\!\left(\tfrac{7}{\varepsilon}\right) \leq D_2,$$

so that

$$N \leq e^{D_2}.$$

Finally,

$$
\begin{aligned}
\frac{\binom{d+m_2}{m_2}}{\binom{d}{m_2}} &= \frac{(d+m_2)!(d-m_2)!}{d!^2} \\
&= \frac{(d+m_2)(d+m_2-1)\cdots(d+1)}{d(d-1)\cdots(d-m_2+1)} \\
&\leq \left(\frac{d+m_2}{d-m_2}\right)^{m_2} = \left(1 + \frac{2m_2}{d-m_2}\right)^{m_2} \leq \left(1 + \frac{4}{m_2}\right)^{m_2} \leq e^4.
\end{aligned}
$$

This concludes the proof of the upper bound:

$$\log N \ \leq \ D_2 \ \lesssim \ \binom{d}{\lceil 4/\varepsilon \rceil^2}.$$

The constant 4 is not optimal and can in fact be reduced essentially to 2.

4.1 Lower bound

For the lower bound, given $\varepsilon > \frac{1}{d^{1/4}}$, our strategy is to begin with a $\frac{1}{2}$-separated set of polynomials of degree at most m, with $m \simeq \frac{1}{\varepsilon^2}$, and from it construct an ε-separated set of Lipschitz functions. By convolving μ with a tiny Gaussian of variance tending to zero, it is not difficult to show that we may assume that μ has and positive density on the whole $\mathbb{R}^n$. From now on we fix such an isotropic product log-concave measure μ and denote by

$$(T_t)_{t \geq 0}$$

the associated Langevin semigroup. One possible way of transforming a polynomial P into a Lipschitz function f_P is to set

$$f_P = T_t(P|_\lambda), \tag{82}$$

for some $\lambda, t > 0$, where $P|_\lambda$ denotes the truncation

$$P|_\lambda(x) = P(x)\, 1_{\{|P(x)| \leq \lambda\}}.$$

By construction $P|_\lambda$ is bounded by λ, thus by Lemma 2.7,

$$f_P \text{ is } \tfrac{\lambda}{\sqrt{t}}\text{-Lipschitz.}$$

We shall choose t and λ so that the L^2 norm of f_P is not too different from that of P. More precisely, we would like to ensure that if P and Q are two polynomials of degree at most m such that

$$\|P - Q\|_{L^2(\mu)} \geq \tfrac{1}{2},$$

then

$$\|f_P - f_Q\|_{L^2(\mu)} \geq c > 0. \tag{83}$$

If we can ensure (83) for any pair of polynomials P, Q in a $\tfrac{1}{2}$-separated set of $\mathcal{P}_{d,m}$, then we will have constructed a c-separated set of Lipschitz functions with Lipschitz constant $\tfrac{\lambda}{\sqrt{t}}$. Equivalently, a $\tfrac{c\sqrt{t}}{\lambda}$-separated set of 1-Lipschitz functions.

Let us discuss what values of t and λ might ensure (83). At this heuristic level, it is helpful to consider the case $\mu = \gamma$. In this case, the Langevin semigroup is the Ornstein–Uhlenbeck semigroup, which acts diagonally on Hermite polynomials:

$$T_t H_\alpha = e^{-t|\alpha|} H_\alpha.$$

Thus, by decomposing a polynomial P of degree at most m into the orthonormal Hermite basis,

$$P = \sum_{|\alpha| \leq m} P_\alpha H_\alpha,$$

we obtain

$$\|T_t P\|_{L^2(\gamma)}^2 = \sum_{|\alpha| \leq m} e^{-2t|\alpha|} P_\alpha^2 \geq e^{-2tm} \sum_{|\alpha| \leq m} P_\alpha^2 = e^{-2tm} \|P\|_{L^2(\gamma)}^2. \tag{84}$$

Although we will apply T_t to the truncated polynomial $P|_\lambda$ rather than to P itself, using the fact that T_t is a contraction in $L^2(\gamma)$ we may write

$$\|T_t(P|_\lambda)\|_{L^2(\gamma)} \geq \|T_t P\|_{L^2(\gamma)} - \|T_t(P - P|_\lambda)\|_{L^2(\gamma)} \geq e^{-tm} \|P\|_{L^2(\gamma)} - \|P - P|_\lambda\|_{L^2(\gamma)}. \tag{85}$$

Thus, if we choose t of order $1/m$, we must choose λ large enough so that

$$\|P - P|_\lambda\|_{L^2(\gamma)}$$

is sufficiently small. The issue is that for an arbitrary degree-m polynomial P with

$$\|P\|_{L^2(\gamma)} = 1,$$

if one wants to truncate at some level $\lambda > 0$ so that

$$\|P - P|_\lambda\|_{L^2(\gamma)} \leq \tfrac{1}{10},$$

one may have to take
$$\lambda \geq e^{cm},$$
which is too large for our purposes. Indeed, the fourth moment of P may be as large as
$$\mathbb{E}P^4(G) \geq e^{cm}$$
for some constant $c > 0$. This can already be seen in dimension one by considering the degree-m monomial.

We resolve this issue by considering random degree-m polynomials, for which we show that, with positive probability, it suffices to take
$$\lambda = \lambda_0 > 0,$$
a constant independent of m. Moving away from the Gaussian setting, we also show that for such random polynomials the Langevin semigroup does not "kill" the L^2 norm too quickly.

4.1.1 Random multilinear polynomials

We restrict our attention to polynomials which are linear combinations of degree-m multilinear monomials. Let
$$D_0 = \binom{d}{m}.$$
Write $\mathcal{S}_{d,m}$ for the collection of all subsets of $\{1, \ldots, d\}$ of cardinality m. We define a polynomial
$$P_\theta = \sum_{\alpha \in \mathcal{S}_{d,m}} \theta_\alpha \prod_{i \in \alpha} X_i = \sum_{\alpha \in \mathcal{S}_{d,m}} \theta_\alpha X_\alpha, \tag{86}$$
for a given vector $\theta \in \mathbb{R}^{D_0} \cong \mathbb{R}^{\mathcal{S}_{d,m}}$, where we write
$$X_\alpha = \prod_{i \in \alpha} X_i.$$

Our intuition is that for a random θ, the value distribution of $P_\theta(X)$ should be roughly Gaussian.

Lemma 4.5. *Assume that $m^2 \leq d$ and let P_θ be defined by (86). Let $\theta \in \mathbb{R}^{D_0}$ be a Gaussian random vector of mean zero and covariance $(1/D_0) \cdot I_{D_0}$. Then the expected fourth moment of P_θ is bounded by*
$$\mathbb{E}_\theta \left[\mathbb{E} P_\theta^4(X) \right] \leq 3 + \frac{Cm^2}{d} \leq C_0, \tag{87}$$
where $X \sim \mu$, C is a universal constant, and $C_0 = 3 + C$.

Proof. We expand
$$\begin{aligned}
\mathbb{E}_\theta \left[\mathbb{E} P_\theta^4(X) \right] &= \mathbb{E} \sum_{\alpha_1, \alpha_2, \alpha_3, \alpha_4} \theta_{\alpha_1} \theta_{\alpha_2} \theta_{\alpha_3} \theta_{\alpha_4} X_{\alpha_1} X_{\alpha_2} X_{\alpha_3} X_{\alpha_4} \\
&= \sum_\alpha \mathbb{E}[\theta_\alpha^4] \, \mathbb{E}[X_\alpha^4] + 3 \sum_{\alpha \neq \beta} \mathbb{E}[\theta_\alpha^2] \, \mathbb{E}[\theta_\beta^2] \, \mathbb{E}[X_\alpha^2 X_\beta^2] \\
&= \frac{3}{D_0^2} \sum_\alpha \mathbb{E}[X_\alpha^4] + \frac{3}{D_0^2} \sum_{\alpha \neq \beta} \mathbb{E}[X_\alpha^2 X_\beta^2] \\
&= \frac{3}{D_0^2} \sum_{\alpha, \beta} \mathbb{E}[X_\alpha^2 X_\beta^2].
\end{aligned}$$

Here we used that
$$\mathbb{E}[\theta_{\alpha_1}\theta_{\alpha_2}\theta_{\alpha_3}\theta_{\alpha_4}] \neq 0$$
if and only if all four indices are equal (giving the first term), or if they form two distinct pairs (three such pairings, giving the second term). Now, for $\alpha, \beta \in \mathcal{S}_{d,m}$,

$$\mathbb{E}\big[X_\alpha^2 X_\beta^2\big] = \mathbb{E}\left(\prod_{i\in\alpha\cap\beta} X_i^4 \prod_{i\in\alpha\cup\beta\backslash(\alpha\cap\beta)} X_i^2\right) \leq 9^{|\alpha\cap\beta|},$$

using that for any centered log-concave random variable X,

$$\mathbb{E}X^4 \leq 9\,\mathbb{E}X^2,$$

see e.g. [Eit24, Theorem 1.4]. Thus

$$
\begin{aligned}
\frac{3}{D_0^2}\sum_{\alpha,\beta}\mathbb{E}[X_\alpha^2 X_\beta^2] &\leq \frac{3}{D_0^2}\sum_{\alpha,\beta} 9^{|\alpha\cap\beta|} \\
&= 3\,\mathbb{E}9^{|\alpha\cap\beta|} \\
&= 3\,\mathbb{E}9^{|\alpha\cap\{1,\dots,m\}|},
\end{aligned}
$$

where in the last line we denote by α and β two independent uniform random subsets of $\{1,\dots,d\}$ of size m, and used invariance under any bijection. The random variable $|\alpha \cap \{1,\dots,m\}|$ follows a hypergeometric distribution:

$$|\alpha \cap \{1,\dots,m\}| \sim \text{Hypergeometric}(d,m,m).$$

It is well known that $\text{Hypergeometric}(N,K,n)$ is stochastically dominated by $\text{Binomial}(n, K/N)$ (and Hoeffding [Hoe63] even proved that the same domination also holds in the convex order). Thus, for any increasing or convex f,
$$\mathbb{E}f\big(|\alpha \cap [1,m]|\big) \leq \mathbb{E}f(Z),$$
where $Z \sim \text{Binomial}(m, p = m/d)$. In particular,

$$
\begin{aligned}
\mathbb{E}9^{|\alpha\cap[1,m]|} \leq \mathbb{E}9^Z &= (1+8p)^m \\
&= \left(1 + \frac{8m}{d}\right)^m \\
&\leq \exp(8m^2/d) \ \leq \ 1 + C_0 m^2/d,
\end{aligned}
$$

where one may take $C_0 = e^8 - 1$. This concludes the proof of Lemma 4.5. $\qquad\square$

We have established that, on average, the 4-th moment of the random multilinear polynomials is bounded. We now need an argument to show that their L^2 norm does not decay too quickly along the Langevin semigroup. We will use Lemma 2.6 from Section 2, which states that for any $f \in L^2(\mu)$ with square integrable gradient,

$$\|T_t^\mu f\|_{L^2(\mu)}^2 \ \geq \ \|f\|_{L^2(\mu)}^2 - 2t\int \|\nabla f\|^2\,d\mu. \tag{88}$$

As before we denote the multilinear polynomials by

$$X_\alpha = \prod_{i\in\alpha} X_i.$$

Clearly, for any $1 \leq i \leq d$

$$\partial_i X_\alpha = \begin{cases} 0, & \text{if } i \notin \alpha, \\ X_{\alpha \setminus \{i\}}, & \text{otherwise.} \end{cases}$$

In particular, since μ is a product measure, for a fixed i, the family $(\partial_i X_\alpha)$ is orthonormal in $L^2(\mu)$. For any $\theta \in \mathbb{R}^{D_0}$,

$$\begin{aligned} \int \|\nabla P_\theta\|^2 d\mu &= \int \sum_{i=1}^d (\partial_i P_\theta)^2 d\mu \\ &= \sum_{i=1}^d \sum_{\substack{\alpha \subset \{1,\ldots,d\} \\ |\alpha|=m}} \int \theta_\alpha^2 (\partial_i X_\alpha)^2 d\mu \\ &= \sum_{\substack{\alpha \subset \{1,\ldots,d\} \\ |\alpha|=m}} \sum_{i=1}^d \theta_\alpha^2 \mathbb{1}_{i \in \alpha} \\ &= m \|\theta\|_2^2 \end{aligned} \tag{89}$$

We are now in a position to prove the lower bound of Theorem 4.1. Let N be an integer to be chosen later, and let $\theta_1, \ldots, \theta_N$ be i.i.d random vectors with distribution

$$\theta_i \sim \mathcal{N}\left(0, \tfrac{1}{D_0} I_{D_0}\right).$$

Let $P_{\theta_1}, \ldots, P_{\theta_N}$ be the corresponding polynomials defined by (86). For any $1 \leq i, j \leq N$ we have

$$\|P_i\|_{L^2(\mu)} = \|\theta_i\|_2 \quad \text{and} \quad \|P_i - P_j\|_{L^2(\mu)} = \|\theta_i - \theta_j\|_2.$$

Furthermore, for any $i \neq j$, the random vector $\theta_i - \theta_j$ is again Gaussian with covariance $\tfrac{2}{D_0} I_{D_0}$. By Gaussian concentration for Lipschitz functions and a union bound, we obtain

$$\begin{aligned} \mathbb{P}(\exists 1 \leq i \neq j \leq N : \|\theta_i - \theta_j\|_2 \leq 1) &\leq N^2 \, \mathbb{P}\left(\|\tfrac{\sqrt{2}}{\sqrt{D_0}} G\|_2 \leq 1\right) \\ &\leq N^2 \, \mathbb{P}\left(\|G\| \leq \mathbb{E}\|G\| - \tfrac{\sqrt{D_0}}{4}\right) \\ &\leq N^2 e^{-D_0/32}, \end{aligned}$$

where $G \sim \mathcal{N}(0, I_{D_0})$, and where we used that

$$\sqrt{D_0} - 1 \ \leq \ \mathbb{E}\|G\|_2.$$

We also have the tail bound

$$\mathbb{P}\left(\exists 1 \leq i \leq N : \|\theta_i\|_2^2 \geq 2\right) \ \leq \ N e^{-D_0/4}.$$

We choose $N = e^{D_0/128}$, and define the events

$$\mathcal{A} = \{(P_{\theta_1}, \ldots, P_{\theta_N}) \text{ is a 1-separated set in } L^2(\mu)\},$$

and

$$\mathcal{B} = \{\|\theta_i\|_2^2 \leq 2 \qquad \forall 1 \leq i \leq N\}.$$

From the two previous inequalities we deduce that

$$\mathbb{P}(\mathcal{A} \cap \mathcal{B}) \geq 1 - 2e^{-D_0/64} \ \geq \ \tfrac{3}{5},$$

38

say. On the other hand, by Lemma 4.5 and Markov's inequality, we have that

$$p = \mathbb{P}\big(\mathbb{E}_X P_\theta^4(X) \leq 2C_0\big) \geq 1/2,$$

where C_0 is the constant from Lemma 4.5. Thus, roughly half of the $(P_{\theta_i})_{1 \leq i \leq N}$ will enjoy a nice bound on their fourth moment. That is, define

$$N_1 = \#\{1 \leq i \leq N : \ \mathbb{E}_X P_{\theta_i}^4(X) \leq 2C_0\} \sim \text{Binomial}(N, p).$$

The median of a Binomial with parameters (N, p) is greater than $\lfloor Np \rfloor$. Thus, with probability $1/2$, we have

$$N_1 \geq \lfloor pN \rfloor \geq N/3.$$

The event

$$\mathcal{D} = \mathcal{A} \cap \mathcal{B} \cap \{N_1 \geq N/3\}$$

has positive probability, greater than 0.1. For such a realization we find polynomials

$$(P_i)_{1 \leq i \leq N_1}$$

that form a 1-separated set of $B_{L^2(\mu)}(0, \sqrt{2})$ of cardinality

$$N_1 \geq N/3 = e^{D_0/128}/3 \geq e^{D_0/256},$$

and satisfy

$$\mathbb{E}_X P_i^4(X) \leq 2C_0 \quad \forall 1 \leq i \leq N_1, \tag{90}$$

$$\|P_i\|_{L^2(\mu)}^2 = \|\theta_i\|_2^2 \leq 2, \tag{91}$$

$$\|\nabla P_i\|_{L^2(\mu)}^2 = m\|\theta_i\|_2^2 \leq 2m. \tag{92}$$

As described above, we set

$$f_i = T_t(P_i|_\lambda) \tag{93}$$

with

$$t = \tfrac{1}{32m},$$

and λ to be chosen later. Then f_i is Lipschitz with constant

$$\frac{\lambda}{\sqrt{t}} = 4\sqrt{2}\lambda\sqrt{m}. \tag{94}$$

Secondly, T_t is a contraction in $L^2(\mu)$, so

$$\|f_i\|_{L^2(\mu)} \leq \|P_i|_\lambda\|_{L^2(\mu)} \leq \|P_i\|_{L^2(\mu)} \leq 2.$$

Let us verify that $(f_i)_{1 \leq i \leq N_1}$ is separated. Let $i \neq j$, using the triangle inequality, (88), (89) and (92), we get

$$\begin{aligned}
\|f_i - f_j\|_{L^2(\mu)} &= \|T_t(P_i|_\lambda - P_j|_\lambda)\|_{L^2(\mu)} \\
&\geq \|T_t(P_i - P_j)\|_{L^2(\mu)} - \|T_t(P_i|_\lambda - P_j|_\lambda - P_i + P_j)\|_{L^2(\mu)} \\
&\geq \left(\|P_i - P_j\|_{L^2(\mu)}^2 - 2t\int \|\nabla(P_i - P_j)\|^2 d\mu\right)^{1/2} - \|P_i - P_i|_\lambda\|_{L^2(\mu)} - \|P_j - P_j|_\lambda\|_{L^2(\mu)} \\
&\geq \left(1 - 4t\big(\|\nabla P_i\|_{L^2(\mu)}^2 + \|\nabla P_j\|_{L^2(\mu)}^2\big)\right)^{1/2} - \|P_i - P_i|_\lambda\|_{L^2(\mu)} - \|P_j - P_j|_\lambda\|_{L^2(\mu)} \\
&\geq (1 - 16tm)^{1/2} - \|P_i - P_i|_\lambda\|_{L^2(\mu)} - \|P_j - P_j|_\lambda\|_{L^2(\mu)} \\
&\geq \tfrac{1}{\sqrt{2}} - \|P_i - P_i|_\lambda\|_{L^2(\mu)} - \|P_j - P_j|_\lambda\|_{L^2(\mu)}.
\end{aligned} \tag{95}$$

It remains to upper-bound

$$\|P_i - P_i|_\lambda\|_{L^2(\mu)} \qquad \forall 1 \le i \le N_1.$$

We fix some $1 \le i \le N$, let

$$\mathcal{E} = \{|P_i| \ge \lambda\}.$$

Using Markov's inequality and (91),

$$\mathbb{P}(\mathcal{E}) \le \frac{\|P_i\|_{L^2(\mu)}^2}{\lambda^2} \le \frac{2}{\lambda^2}. \tag{96}$$

Using Cauchy-Schwarz and (90),

$$
\begin{aligned}
\|P_i - P_i|_\lambda\|_{L^2(\mu)} &= \|P_i \mathbb{1}_{\mathcal{E}}\|_{L^2(\mu)} \\
&\le \|P_i\|_{L_4(\mu)} \|\mathbb{1}_A\|_{L_4(\mu)} \\
&\le \left(\frac{2C_0}{\lambda^2}\right)^{1/4}.
\end{aligned}
$$

We choose

$$\lambda = 16\sqrt{2}\sqrt{C_0}$$

and we find that for all $1 \le i \le N_1$,

$$\|P_i - P_i|_\lambda\|_{L^2(\mu)} \le \frac{1}{4}.$$

Plugging this back into (95), we arrive at

$$\|f_i - f_j\|_{L^2(\mu)} \ge \frac{1}{\sqrt{2}} - \frac{1}{2} \ge \frac{1}{5} \tag{97}$$

for all $1 \le i \ne j \le N_1$. Setting

$$\tilde{f}_i = \frac{1}{4\sqrt{2}\sqrt{m}\lambda} f_i = \frac{1}{128\sqrt{C_0}\sqrt{m}} f_i$$

We have constructed a family of 1-Lipschitz functions which is $\frac{\tilde{C}}{\sqrt{m}}$-separated and has cardinality

$$N_1 \ge e^{D_0/256}$$

where

$$D_0 = \binom{d}{m}.$$

In other words, for a given $\varepsilon > 0$, setting

$$m = \lfloor \frac{\tilde{C}^2}{\varepsilon^2} \rfloor$$

we constructed an ε-separated set of cardinality of cardinality N_1 with

$$\log N_1 \gtrsim \binom{d}{m} \ge \binom{d}{\frac{c}{\varepsilon^2}}$$

for some constant $c > 0$, which is what we wanted to prove.

4.1.2 The general case

We explain how to deduce the general isotropic case from the case of the product measure. It is well-known that lower-dimensional marginals of an isotropic log-concave probability measure are approximately Gaussian. The following precise statement was proved in [EK08]:

Theorem 4.6. *Let μ be an isotropic log-concave probability measure in $\mathbb{R}^d$. Then there exists a subspace $E \subseteq \mathbb{R}^d$ of dimension $k \geq d^{\eta_0}$ such that*

$$|p_E(x) - q_{\gamma_E}(x)| \leq \frac{C}{k} q_{\gamma_E}(x) \qquad \text{for all } |x| \leq k \tag{98}$$

where C, η_0 are universal constants, p_E is the density of the marginal μ_E of μ on E and $q_{\gamma_E}(x) = (2\pi)^{-k/2} e^{-|x|^2/2}$ is the density of a standard Gaussian on E, which we denote by γ_E.

The estimate (98) implies that p_E is very close to γ_E on a ball of radius k, while most of the mass of μ_E, or γ_E, is concentrated in a ball of radius only $\simeq \sqrt{k}$. This implies in particular that the L^2 norm of a Lipschitz function does not change much when swithcing from μ_E to γ_E. Indeed, let g be a 1-Lipschitz function. Then,

$$\int_E g^2 d\mu \geq \int_{|x| \leq k} g^2 d\mu$$

$$= \int g^2 d\gamma_E - \int_{|x| \leq k} g^2(d\gamma_E - d\mu_E) - \int_{|x| > k} g^2 d\gamma_E$$

$$\geq \int g^2 d\gamma_E - \frac{C}{k} \int g^2 d\gamma_E - \left(\int g^4 d\gamma_E\right)^{1/2} \mathbb{P}\left(|G_E| \geq k\right)^{1/2}$$

$$\geq \int g^2 d\gamma_E \left(1 - \frac{C}{k}\right) - C_1 e^{-k} \tag{99}$$

where C_1 is a universal constant. In the last line we have used concentration of the norm of a standard k-dimensional Gaussian:

$$\gamma_k\left(|x| \geq \sqrt{k} + t\right) \leq e^{-t^2/2}.$$

and that for all 1-Lipschitz functions g,

$$\left(\int g^4 d\gamma\right)^{1/2} \leq \int g^2 d\gamma + \tilde{C}_1$$

for some constant $\tilde{C}_1$.

Let $1 > \varepsilon > \frac{1}{k^{1/4}}$. By Theorem 4.1, we can find 1-Lipschitz functions $f_1, \ldots, f_N$ such that for all $i \neq j$

$$\|f_i - f_j\|_{L^2(\gamma_E)} \geq \varepsilon$$

and

$$\log N \gtrsim k^{\frac{c}{\varepsilon^2}} \gtrsim d^{\frac{c'}{\varepsilon^2}}.$$

We now apply (99) to the 1-Lipschitz function $g = \frac{1}{2}(f_i - f_j)$. For large enough d, we get

$$\|f_i - f_j\|_{L^2(\mu_E)}^2 \geq \|f_i - f_j\|_{L^2(\gamma_E)}^2 \left(1 - \frac{C}{k}\right) - 4C_1 e^{-k}$$

$$\geq \varepsilon^2 \left(1 - \frac{C}{k}\right) - 4C_1 e^{-k}$$

$$\geq \frac{\varepsilon^2}{4}$$

where we assume that d is large enough so that $\frac{C}{k} \leq \frac{1}{2}$ and $4C_1 e^{-k} \leq \frac{1}{4\sqrt{k}} \leq \frac{\varepsilon^2}{4}$. Thus the functions $f_1, \ldots, f_N$ are $\varepsilon/2$ separated in $L^2(\mu_E)$. Equivalently, the functions $f_1 \circ \Pi_E, \ldots, f_N \circ \Pi_E$ are $\varepsilon/2$ separated in $L^2(\mu)$, where $\Pi_E : \mathbb{R}^d \to E$ is the orthogonal projection operator. Thus for all $\varepsilon > d^{-\eta_0/4}$,

$$H_L^\mu(\varepsilon) \geq \log N \gtrsim d^{\frac{c'}{\varepsilon^2}}.$$

This proves the general case, with $\eta = \eta_0/4$.

4.2 A Minimax lower bound for learning Lipschitz functions

We now go back to the learning problem

$$Y_i = f(X_i) + \sigma Z_i \quad i = 1, \ldots, n. \tag{100}$$

and we prove the minimax lower bound announced in the Introduction, Corollary 1.6, which we restate below for the reader's convenience.

Corollary (Corollary 1.6). *Let μ be an isotropic log-concave measure on $\mathbb{R}^d$. Assume that the noise satisfies*

$$n^{-\kappa} \leq \sigma^2 \leq n$$

for some constant $\kappa > 0$. There exists a universal constant $c > 0$ such that if

$$n \leq e^{\frac{cd^{2\eta} \log d}{\kappa}},$$

the minimax risk is lower bounded as

$$\mathcal{R}_{n,d}^* \gtrsim (1 + \kappa) \frac{\log n}{\log d}. \tag{101}$$

Moreover, if additionally μ is a product measure, then the lower bound (101) *holds in the range*

$$n \leq e^{\frac{c\sqrt{d} \log d}{\kappa}}.$$

A typical way of establishing lower bounds for a learning problem is to reduce it to a multiple hypothesis testing problem and apply information-theoretic methods. This is known as Fano's method. More precisely, we shall use the Yang–Barron version, which requires computing entropy estimates in the Kullback–Leibler divergence for the collection of random variables

$$\mathcal{D}_n = \{((X_1, Y_1), \ldots, (X_n, Y_n)) : f \in B_2(\mathrm{Lip}, 0, 1)\} = \{(X, Y_f) : f \in B_2(\mathrm{Lip}, 0, 1)\},$$

namely, the collection of all possible random vectors that we may observe, indexed by our function class, the 1-Lipschitz functions with bounded $L^2(\mu)$ norm. Let P and Q be two probability measures on $\mathbb{R}^d$ such that P is absolutely continuous with respect to Q. The Kullback–Leibler divergence from P to Q, denoted by $\mathrm{D_{KL}}(P \,\|\, Q)$, is defined as

$$\mathrm{D_{KL}}(P \,\|\, Q) := \int_{\mathbb{R}^d} \log\left(\frac{dP}{dQ}\right) dP,$$

where $\frac{dP}{dQ}$ denotes the Radon–Nikodym derivative of P with respect to Q. For $\varepsilon > 0$, let

$$\tilde{\mathcal{N}}(\mathcal{D}_n, \varepsilon, \mathrm{D_{KL}})$$

be the minimal size of an ε-net of $\mathcal{D}_n$ with respect to $\mathrm{D_{KL}}$, and set

$$\tilde{\mathrm{H}}(\mathcal{D}_n, \varepsilon, \mathrm{D_{KL}}) := \log \tilde{\mathcal{N}}(\mathcal{D}_n, \varepsilon, \mathrm{D_{KL}}).$$

Here the entropy is defined via covering numbers; we use a tilde to emphasize the distinction from the earlier convention adopted for H_L, which was based on packing numbers.

The Yang–Barron method is summarized in the next lemma; see [Wai19] for background.

Lemma 4.7. *Let $\varepsilon > 0$ be such that*

$$\varepsilon^2 \geq \tilde{H}(\mathcal{D}_n, \varepsilon, D_{\mathrm{KL}}), \tag{102}$$

and $\delta > 0$ be such that

$$H_L(\delta) \geq 4\varepsilon^2 + 2\log 2. \tag{103}$$

Then, the minimax risk using n samples is lower bounded as

$$\inf_{\hat{f}} \sup_{f \in B_2(\mathrm{Lip},0,1)} \mathbb{E}\|f - \hat{f}\|_{L^2(\mu)}^2 \gtrsim \delta^2. \tag{104}$$

Proof of Corollary 1.6. We first compute the metric entropy of $\mathcal{D}_n$ equipped with the Kullback–Leibler divergence. Let f_1 and f_2 be two Lipschitz functions. For $k = 1, 2$, the vector Y_{f_k} decomposes as

$$Y_{f_k} = f_k(X) + G_k,$$

where $f_k(X) = (f_k(X_1), \ldots, f_k(X_n))$ and $G_k \sim \mathcal{N}(0, \sigma^2 I_n)$ are independent Gaussian vectors. Conditioning on X, Y_{f_1} and Y_{f_2} are independent Gaussians with means $f_1(X)$ and $f_2(X)$, and covariance $\sigma^2 I_n$. It follows that

$$
\begin{aligned}
D_{\mathrm{KL}}((X, Y_{f_1}) \,\|\, (X, Y_{f_2})) &= \mathbb{E}\big[D_{\mathrm{KL}}\big(Y_{f_1} \mid X \,\|\, Y_{f_2} \mid X \big) \big] \\
&= \mathbb{E}\left(\frac{1}{2\sigma^2} \sum_{i=1}^{n} \big(f_1(X_i) - f_2(X_i) \big)^2 \right) \\
&= \frac{n}{2\sigma^2} \|f_1 - f_2\|_{L^2(\mu)}^2.
\end{aligned}
$$

In particular, choosing $f_1 = 0$, the radius of $\mathcal{D}_n$ in Kullback–Leibler divergence is at most

$$\frac{n}{2\sigma^2}.$$

Set

$$\varepsilon^2 = \frac{n}{2\sigma^2},$$

which trivially ensures (102). To satisfy (103), we require

$$H_L(\delta) \gtrsim \frac{n}{\sigma^2}.$$

By Theorem 4.1, provided that $\delta \geq d^{-\eta}$ in the general case (respectively, $\delta \geq d^{-1/4}$ in the product case), it suffices that

$$d^{c/\delta^2} \geq \frac{n}{\sigma^2},$$

i.e.

$$\delta^2 \lesssim \frac{\log d}{\log(n/\sigma^2)}.$$

Using $\sigma^2 \geq n^{-\kappa}$, we have $\log(n/\sigma^2) \geq (1 + \kappa) \log n$, so we may take

$$\delta^2 = \frac{c}{1 + \kappa} \frac{\log d}{\log n}$$

for some $c > 0$. The applicability condition $\delta \geq d^{-\eta}$ (respectively $\delta \geq d^{-1/4}$) amounts to

$$\delta^2 \geq d^{-2\eta} \iff \frac{c}{1 + \kappa} \frac{\log d}{\log n} \geq d^{-2\eta},$$

(respectively $\frac{c}{1+\kappa} \frac{\log d}{\log n} \geq d^{-1/2}$) which holds whenever

$$n \leq \exp\left(\frac{c\, d^{2\eta}\, \log d}{1+\kappa}\right).$$

Respectively,

$$n \leq \exp\left(\frac{c\, \sqrt{d}\, \log d}{1+\kappa}\right).$$

This yields the stated bounds. $\qquad\square$

References

[AS17]   Guillaume Aubrun and Stanisław J Szarek. *Alice and Bob meet Banach*, volume 223. American Mathematical Soc., 2017.

[BGL13] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

[Biz23] Pierre Bizeul. On the log-sobolev constant of log-concave measures. *arXiv preprint arXiv:2306.12997*, 2023.

[BK25]   Pierre Bizeul and Boaz Klartag. Polynomial approximation in l^2 of the double exponential via complex analysis. *arXiv preprint arXiv:2502.07448*, 2025.

[Bob03] Sergey G Bobkov. On concentration of distributions of random weighted sums. *Annals of probability*, pages 195–215, 2003.

[Bou91] Jean Bourgain. On the distribution of polynomials on high-dimensional convex sets. In *Geometric aspects of functional analysis (1989–90)*, volume 1469 of *Lecture Notes in Math.*, pages 127–137. Springer, Berlin, 1991.

[CW01]   Anthony Carbery and James Wright. Distributional and lq norm inequalities for polynomials over convex bodies in rn. *Mathematical research letters*, 8(3):233–248, 2001.

[EI22]   Alexandros Eskenazis and Paata Ivanisvili. Learning low-degree functions from a logarithmic number of random queries. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 203–207, 2022.

[EIS22] Alexandros Eskenazis, Paata Ivanisvili, and Lauritz Streck. Low-degree learning and the metric entropy of polynomials. *arXiv preprint arXiv:2203.09659*, 2022.

[Eit24] Yam Eitan. The centered convex body whose marginals have the heaviest tails. *Studia Math.*, 274(3):201–215, 2024.

[EK08]   Ronon Eldan and Boaz Klartag. Pointwise estimates for marginals of convex bodies. *J. Funct. Anal.*, 254(8):2275–2293, 2008.

[Fre77] Géza Freud. On markov-bernstein-type inequalities and their applications. *Journal of Approximation Theory*, 19(1):22–37, 1977.

[GM83]   Mikhael Gromov and Vitali D Milman. A topological application of the isoperimetric inequality. *American Journal of Mathematics*, 105(4):843–854, 1983.

[Hoe63]   Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[KL25]   Boaz Klartag and Jospeh Lehec. Isoperimetric inequalities in high dimensional convex sets. *Bulletin of the Amer. Math. Soc.*, 2025+.

[Kla23]   Boaz Klartag. Logarithmic bounds for isoperimetry and slices of convex sets. *arXiv preprint arXiv:2303.14938*, 2023.

[KLS95]   Ravi Kannan, László Lovász, and Miklós Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13:541–559, 1995.

[LL87]   A.L. Levin and D. S. Lubinsky. Canonical products and the weights $\exp(-x^\alpha)$ $(\alpha > 1)$ with applications. *Journal of approximation theory*, 49(2):149–169, 1987.

[Lub07]   Doron S Lubinsky. A survey of weighted approximation for exponential weights. *arXiv preprint math/0701099*, 2007.

[Mil09]   Emanuel Milman. On the role of convexity in isoperimetry, spectral gap and concentration. *Inventiones mathematicae*, 177(1):1–43, 2009.

[NSV02]   Fedor Nazarov, Mikhail Sodin, and Alexander Volberg. The geometric Kannan-Lovász-Simonovits lemma, dimension-free estimates for the distribution of the values of polynomials, and the distribution of the zeros of random analytic functions. *Algebra i Analiz*, 14(2):214–234, 2002.

[Pis03]   Gilles Pisier. *Introduction to operator space theory*. Number 294. Cambridge University Press, 2003.

[Rud99]   Mark Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72, 1999.

[Ver18]   Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[Wai19]   Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.