Near-optimality of $\Sigma\Delta$ quantization for L^2-approximation with polynomials in Bernstein form

C. Sinan Güntürk
NYU Courant Institute
New York, NY, USA
gunturk@cims.nyu.edu

Weilin Li
CUNY City College
New York, NY, USA
wli6@ccny.cuny.edu

***Abstract*—In this paper, we provide lower bounds on the L^2-error of approximation of arbitrary functions $f : [0,1] \to \mathbb{R}$ by polynomials of degree at most n, with the constraint that the coefficients of these polynomials in the Bernstein basis of order n are bounded by n^α for some $\alpha \geq 0$. For Lipschitz functions, this lower bound matches, up to a factor of $\sqrt{\log n}$, a previously obtained constructive upper bound for the error of approximation by one-bit polynomials in Bernstein form via $\Sigma\Delta$ quantization where the functions are bounded by 1 and the coefficients of the approximating polynomials are constrained to be in $\{\pm 1\}$.**

I. Introduction and Statement of the Main Theorem

For any natural number n, let $\mathcal{B}_n := (p_{n,k})_{k=0}^n$, where

$$p_{n,k}(x) := \binom{n}{k} x^k (1-x)^{n-k}, \quad x \in [0,1],$$

denote the Bernstein basis of order n for the linear space $\mathcal{P}_n$ of polynomials of degree at most n, considered as a subspace of real-valued functions on $[0,1]$. Consider the "synthesis map" $S_n : \mathbb{R}^{n+1} \to \mathcal{P}_n$ associated with the Bernstein basis:

$$S_n u := \sum_{k=0}^n u_k p_{n,k}, \quad u \in \mathbb{R}^{n+1}. \tag{1}$$

In recent work [1], it was shown that for every continuous function $f : [0,1] \to [-1,1]$ and for every positive integer n, there exists a sign vector $\sigma := (\sigma_0, \dots, \sigma_n) \in \{\pm 1\}^{n+1}$ such that

$$\left| f(x) - (S_n\sigma)(x) \right| \lesssim \omega_f\Big(\frac{1}{\sqrt{n}}\Big) + \min\left(1, \frac{1}{\sqrt{nX}}\right), \tag{2}$$

where ω_f stands for the modulus of continuity of f and $X := x(1-x)$. Here, $A_n \lesssim B_n$ means $A_n \leq CB_n$ for all n where C is an absolute constant. When C depends on some parameter α, we use the notation $\lesssim_\alpha$. In fact, a more refined version of the bound (2) was shown in [1], but this refinement will not be needed in this note.

The sign vector σ is computed constructively, in linear time, from $n+1$ regular samples of f on $[0,1]$ by means of first-order $\Sigma\Delta$ quantization, which is a well-known analog-to-digital conversion method. (See e.g. [5] for theory and engineering applications.) Note that $\sum_k p_{n,k} = 1$ so that $\|S_n\sigma\|_\infty \leq 1$, therefore $\|f\|_\infty \leq 1$ is necessary for approximability.

While the error bound (2) is not uniform in x, it offers p-norm bounds on $[0,1]$ for all $p < \infty$. When $p = 2$ and $f : [0,1] \to [-1,1]$ is Lipschitz, it follows easily that

$$\left\| f - S_n\sigma \right\|_2 \lesssim \frac{|f|_{\mathrm{Lip}}}{\sqrt{n}} + \sqrt{\frac{\log n}{n}}. \tag{3}$$

The $\log n$ term is removable if $\|f\|_\infty < 1$. In this case, for every $\mu < 1$ and $\|f\|_\infty \leq \mu$, it is also shown in [1] that using second order $\Sigma\Delta$ quantization yields

$$\left| f(x) - (S_n\sigma)(x) \right| \lesssim_\mu \omega_f(\frac{1}{\sqrt{n}}) + \min\left(1, \frac{1}{nX}\right), \tag{4}$$

and therefore

$$\left\| f - S_n\sigma \right\|_2 \lesssim_\mu \frac{1 + |f|_{\mathrm{Lip}}}{\sqrt{n}}. \tag{5}$$

It is natural to ask if the $1/\sqrt{n}$ term above is tight in any sense. The ε-capacity of the Lipschitz ball

$$\mathcal{L} := \Big\{ f \in \mathrm{Lip}([0,1]) : \|f\|_\infty \leq 1, \ |f|_{\mathrm{Lip}} \leq 1 \Big\}$$

in $L^p([0,1])$ is the logarithm (base 2) of the maximal number of points that are ε-separated (with respect to $\|\cdot\|_p$ distance) in $\mathcal{L}$. (See [2] as well as [3], [6].) It is known that this number is bounded below (as well as above, up to p-dependent constants) by $1/\varepsilon$, hence the covering radius of any set of N points is at least of order $1/\log_2 N$. In our setting, this means that we cannot expect approximation of general $f \in \mathcal{L}$ by polynomials of the form $S_n\sigma$ with accuracy better than $1/n$. Hence there is a gap, roughly of order $1/\sqrt{n}$ (depending on whether we assume $\mu < 1$ or $\mu = 1$), between the achievable upper bound in the 2-norm and this universal entropic lower bound.

However, using the entropic lower bound ignores the specific constraints of approximation using both one-bit coefficients *and* the Bernstein basis at the same time. How these two constraints interact with each other is to be understood. There is, in fact, a trivial obstruction to achieving high approximation accuracy near the endpoints of $[0,1]$: For any $\sigma \in \{\pm 1\}^{n+1}$, we have $|(S_n\sigma)(0)| = |(S_n\sigma)(1)| = 1$. It can be checked that the derivative satisfies $\|(S_n\sigma)'\|_\infty \le 2n$, therefore we have $|(S_n\sigma)(x)| \ge 1/2$ whenever $\min(x, 1-x) \le 1/(4n)$, implying that $\|S_n\sigma\|_2 \ge 1/\sqrt{8n}$. In other words, it is not possible to approximate $f = 0$ to accuracy of order better than $1/\sqrt{n}$. A similar lower bound applies to any constant function with its value in $(-1, 1)$.

Even if we allowed for non-discrete coefficients, but still in $[-1, 1]$, the set of polynomials that are available for approximation is limited by the choice of the basis. Geometrically, the problem is to understand the degree to which the parallelotope $S_n([-1,1]^{n+1})$, or its vertices given by $S_n(\{\pm 1\}^{n+1})$, can approximate $\mathcal{L}$.

For this purpose, given any $n \in \mathbb{N}$, $f : [0,1] \to \mathbb{R}$, and $U \subset \mathbb{R}^{n+1}$, let us define

$$E_n(f;U)_p := \inf_{u\in U} \left\|f - S_n u\right\|_p, \tag{6}$$

which measures the error of best approximation to f from $\mathcal{P}_n$ in the p-norm with coefficients in the Bernstein basis $\mathcal{B}_n$ chosen from U. For a class of functions $\mathcal{F}$, we define

$$E_n(\mathcal{F};U)_p := \sup_{f\in\mathcal{F}} E_n(f;U)_p. \tag{7}$$

With this notation, the above findings can be summarized as

$$n^{-1/2} \lesssim E_n(\mu\mathcal{L}; \{\pm1\}^{n+1})_2 \lesssim_\mu \begin{cases} n^{-1/2}, & \mu < 1, \\ (n/\log n)^{-1/2}, & \mu = 1. \end{cases}$$

Our result in this paper will show that the constructive upper bound is actually tight up to a factor of $\log n$ even when the coefficients are chosen without discretization from $[-1, 1]$. In other words, as far as $\mathcal{L}$ is concerned, the discreteness of the coefficients can only play a secondary role in influencing the actual behaviour of $E_n(\mathcal{L}, \{\pm1\}^{n+1})_2$. In fact, our main result given in Theorem 1 states that the above lower bound persists over a much wider range of bounded (but otherwise arbitrary) real-valued coefficients:

Theorem 1. *For any $\alpha \ge 0$,*

$$E_n(\mathcal{L}, n^\alpha[-1,1]^{n+1})_2 \gtrsim_\alpha \frac{1}{\sqrt{n\log n}}. \tag{8}$$

This result may seem surprising at first, but it is rooted in the fact that, numerically speaking, the Bernstein basis can only span a $O(\sqrt{n})$ dimensional space effectively. It was shown in [1] that the ϵ-numerical rank of $\mathcal{B}_n$, i.e. the number of singular values s_k of S_n that lie above ϵs_0, is asymptotic to $\sqrt{2n\log(1/\epsilon)}$. More precisely, the singular values s_k undergo a phase transition at $k \approx \sqrt{n}$. When we prove our theorem in the next section, we will make use of the specific distribution of the singular values to quantify this phase transition.

II. The Proof of the Main Theorem

A. Singular values and singular vectors of S_n

It was shown in [1] that S_n has the singular value decomposition

$$S_n u = \sum_{k=0}^{n} s_k \langle u, \varphi_k\rangle \psi_k \tag{9}$$

where $\psi_0, \dots, \psi_n$ are the first $n+1$ continuous Legendre polynomials on $[0,1]$, $\varphi_0, \dots, \varphi_n$ are the discrete Legendre polynomials on $\{0, \dots, n\}$, related by $S_n\varphi_k = s_k\psi_k$, and the singular values $s_k := s_k(n)$ are given (in decreasing order) by

$$s_k = \sqrt{\frac{(n)_k}{(n+k+1)_{k+1}}}, \quad k = 0, \dots, n.$$

Here $(t)_k := t(t-1)\dots(t-k+1)$ denotes the falling factorial function with $(t)_0 = 1$ and $\langle\cdot,\cdot\rangle$ is the Euclidean inner-product on $\mathbb{R}^{n+1}$.

We recall that both families of Legendre polynomials are orthonormal, the former in $L^2([0,1])$ and the latter in $L^2(\{0, \dots, n+1\})$ which is identified with $\mathbb{R}^{n+1}$. It is important to note that the φ_k depend on n, meanwhile the ψ_k do not. By nature of their definition, we have

$$\operatorname{span}(\psi_0, \dots, \psi_k) = \mathcal{P}_k$$

for all k.

It's more convenient to work with the eigenvalues $\lambda_k := s_k^2$ of $S_n^* S_n$. The following is a simple upper bound:

Lemma 2. *The eigenvalues λ_k of $S_n^* S_n$ satisfy*

$$\lambda_k \le \frac{e^{-k^2/(n+k)}}{n+k+1}, \quad k = 0, \dots, n.$$

Proof. We have $1 - x \le e^{-x}$ for all $x \in \mathbb{R}$ so that for all $j \ge 0$ we have

$$\frac{n-j}{n+k-j} \le 1 - \frac{k}{n+k} \le e^{-k/(n+k)}.$$

Using this bound, it follows at once that

$$\begin{aligned} \lambda_k &= \frac{1}{n+k+1}\prod_{j=0}^{k-1}\frac{n-j}{n+k-j} \\ &\le \frac{1}{n+k+1}\exp(-k^2/(n+k)). \end{aligned}$$

□

B. Proof of Theorem 1.

We proceed with the proof of Theorem 1. Suppose we are given any $f \in L^2([0,1])$ and $u \in \mathbb{R}^{n+1}$. For any $m \ge 0$, let $\mathbf{P}_m$ be the orthogonal projection operator onto $\mathcal{P}_m$, which we can express as

$$\mathbf{P}_m f = \sum_{k=0}^{m} \langle f, \psi_k\rangle_{L^2} \psi_k.$$

Since $\mathbf{P}_m f$ is the best L^2-approximation to f from $\mathcal{P}_m$, we have

$$\begin{aligned} \|f - \mathbf{P}_m f\|_2 &\le \|f - \mathbf{P}_m(S_n u)\|_2 \\ &\le \|f - S_n u\|_2 + \|S_n u - \mathbf{P}_m(S_n u)\|_2. \end{aligned} \tag{10}$$

Notice that for $0 \le m \le n-1$

$$S_n u - \mathbf{P}_m(S_n u) = \sum_{k=m+1}^{n} s_k \langle u, \varphi_k\rangle \psi_k$$

so that

$$\|S_n u - \mathbf{P}_m(S_n u)\|_2^2 = \sum_{k=m+1}^{n} s_k^2 |\langle u, \varphi_k \rangle|^2 \le s_{m+1}^2 \|u\|_2^2.$$

Plugging this bound in (10), it follows that

$$\|f - S_n u\|_2 \ge \|f - \mathbf{P}_m f\|_2 - s_{m+1}\|u\|_2. \tag{11}$$

It is important to note that this bound is valid for all $f \in L^2([0,1])$, $u \in \mathbb{R}^{n+1}$, and $0 \le m \le n-1$. Taking the infimum of both sides over $u \in U := n^\alpha[-1,1]^{n+1}$ yields

$$E_n(f, n^\alpha[-1,1]^{n+1})_2 \ge \|f - \mathbf{P}_m f\|_2 - s_{m+1} n^\alpha \sqrt{n+1} \tag{12}$$

and further taking the supremum of both sides over $f \in \mathcal{L}$ yields

$$E_n(\mathcal{L}, n^\alpha[-1,1]^{n+1})_2 \ge \sup_{f \in \mathcal{L}} \|f - \mathbf{P}_m f\|_2 - s_{m+1} n^\alpha \sqrt{n+1}. \tag{13}$$

We note that we are still free to choose m. We first seek a simple lower bound on $\sup_{f\in\mathcal{L}} \|f - \mathbf{P}_m f\|_2$ which will then give us a suitable reference value to optimally choose m. It will suffice to utilize the known bounds concerning the Kolmogorov m-width $d_m(\mathcal{L})_2$ of the Lipschitz ball $\mathcal{L}$ in $L^2([0,1])$ which is defined to be the infimum, over all m-dimensional linear subspaces $X_m \subset L^2([0,1])$, of the deviation of $\mathcal{L}$ from X_m given by

$$\sup_{f\in\mathcal{L}} \inf_{g \in X_m} \|f - g\|_2.$$

It is known (e.g. [3], [4]) that

$$d_m(\mathcal{L})_2 \gtrsim m^{-1}.$$

Hence we can immediately conclude that

$$\sup_{f\in\mathcal{L}} \|f - \mathbf{P}_m f\|_2 \gtrsim d_{m+1}(\mathcal{L})_2 \gtrsim \frac{1}{m+1}. \tag{14}$$

By Lemma 2, we know that

$$s_{m+1} \le \frac{e^{-(m+1)^2/(4n)}}{\sqrt{n+1}},$$

hence injecting this bound and the bound (14) into (13), we obtain

$$E_n(\mathcal{L}, n^\alpha[-1,1]^{n+1})_2 \gtrsim \frac{1}{m+1} - n^\alpha e^{-(m+1)^2/(4n)}. \tag{15}$$

Setting $m+1 = \lceil C\sqrt{n\log n}\rceil$ where $C \ge 2\sqrt{\alpha+1}$, we get that

$$n^\alpha e^{-(m+1)^2/(4n)} \le n^{\alpha - C^2/4} \le n^{-1}$$

so that

$$E_n(\mathcal{L}, n^\alpha[-1,1]^{n+1})_2 \gtrsim_\alpha \frac{1}{\sqrt{n\log n}} \tag{16}$$

for all sufficiently large n. □

III. Extensions and discussion

It is evident from the proof of Theorem 1 in Section II-B that the result is immediately generalizable to other function classes for which the m-widths are known, as the main lower bound (12) is valid for any function. The Lipschitz ball is the same as the class B^1_∞, the unit ball of the Sobolev space $W^1_\infty([0,1])$ defined by absolutely continuous functions with derivative in $L^\infty([0,1])$. Analogously, the class B^r_p is defined via the Sobolev space $W^r_p([0,1])$, $r \in \mathbb{N}_+$. For this class, the m-width $d_m(B^r_p)_2$ is known (see [3], [4]) to be equivalent to m^{-r} with constants that depend on r and p. Hence a similar selection of $m \asymp \sqrt{n\log n}$ as done above yields

$$E_n(B^r_p, n^\alpha[-1,1]^{n+1})_2 \gtrsim_{\alpha,r,p} \frac{1}{(n\log n)^{r/2}}. \tag{17}$$

The method of $\Sigma\Delta$ quantization was also applied to functions of higher regularity in [1], and non-uniform pointwise error bounds that reflect additional smoothness were obtained. For example, it was shown that if $f \in B^r_\infty$ for $r \ge 2$ and $\|f\|_\infty \le \mu < 1$, then there exists $\sigma \in \{\pm 1\}^{n+1}$ such that

$$|f(x) - (S_n\sigma)(x)| \lesssim_{\mu,r} \|f\|_{W^r_\infty} n^{-r/2} + \min(1, X^{-r} n^{-r/2}) \tag{18}$$

for all $n \gtrsim \|f^{(2)}\|_\infty/(1-\mu)$. While no method can produce high accuracy over all of $[0,1]$ due to the constraints near the endpoints, these pointwise upper bounds provide much faster decay guarantees for the approximation error on compact subintervals $[\delta, 1-\delta] \subset (0,1)$.

Returning to lower bounds, the methods and analysis in this paper relied significantly on the 2-norm. We leave the case of other p-norms for future work.

References

[1] C. Sinan Güntürk and Weilin Li. Approximation with one-bit polynomials in Bernstein form. *Constructive Approximation*, 57(2):601–630, 2023.

[2] A. N. Kolmogorov and V. M. Tihomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspehi Mat. Nauk.*, 14(2 (86)):3–86, 1959. Also in Amer. Math. Soc. Transl., Ser. 2 17 (1961), 277–364.

[3] George G Lorentz, Manfred v Golitschek, and Yuly Makovoz. *Constructive approximation: advanced problems*, volume 304. Springer, 1996.

[4] Allan Pinkus. *N-widths in Approximation Theory*, volume 7. Springer Science & Business Media, 2012.

[5] Richard. Schreier and Gabor. C. Temes. *Understanding Delta-Sigma Data Converters*. Wiley-IEEE Press, 2004.

[6] Anatoliĭ Georgievich Vitushkin. *Theory of the Transmission and Processing of Information*. Pergamon Press, 1961.