The quest for optimal sampling strategies for learning sparse approximations in high dimensions

Juan M. Cardenas
Department of Mathematics
jcardena@sfu.ca

Nick Dexter
Department of Mathematics
nicholas_dexter@sfu.ca

Sebastian Moraga
Department of Mathematics
sebastian_moraga_scheuermann@sfu.ca

Ben Adcock
Department of Mathematics
ben_adcock@sfu.ca

***Abstract*—Learning an accurate approximation to an unknown function from data is a fundamental problem at the heart of many key tasks in computational science and engineering. It presents various challenges, including: the *curse of dimensionality*, which renders classical approaches poorly suited; the fact that obtaining samples is expensive; the potential for the domain to be irregular; and the fact that the target function may take values in an infinite-dimensional Hilbert space. A highly fruitful way to strive to overcome these challenges is by exploiting the fact that functions arising in such applications often admit approximately *sparse* representations in a given dictionary. Accordingly, the purpose of this work is to examine the following question: *supposing multivariate function has an approximately sparse representation, how many samples suffice to learn such an approximation from data, and how can it be computed?* We focus on two scenarios. First, when the representing dictionary elements are known, with the problem being solved via least squares, and second the substantially more challenging scenario where such elements are unknown. We address this using ℓ^1-minimization strategies. Our results apply to scalar- and Hilbert-valued functions. We also introduce a novel ℓ^1-minimization strategy for sparse approximation on irregular domains.**

I. Introduction

Let $(D, \mathcal{D}, \rho)$ be a probability space, where $D \subseteq \mathbb{R}^d$, and $\mathbb{V}$ be a Hilbert space. We consider the problem of approximating a function $f : D \to \mathbb{V}$ from noisy evaluations of f at m *sample points* $y_1, \ldots, y_m$. Our focus is on designing sampling strategies that are *sample efficient*. To this end, we assume that the y_i are independent, with $y_i \sim \mu_i$ for some probability measure μ_i on D. In what follows, we are particularly interested in whether or not standard Monte Carlo (MC) sampling, i.e. $\mu_i = \rho$, $\forall i$, leads to optimal sample complexity bounds. Given samples y_i, we consider data of the form

$$b_i = f(y_i) + n_i \in \mathbb{V}_h, \quad i = 1, \ldots, m.$$

Here $\mathbb{V}_h$ is a finite-dimensional discretization of $\mathbb{V}$ (e.g. it may be a finite element space when f represents a solution of a parametric PDE). We assume such a space is available in what follows.

Next, we consider a dictionary of scalar-valued functions $\Phi = \{\phi_l : l \in \mathcal{I}\} \subset L^2_\rho(D)$, which may be finite, countable or uncountable, and we assume that f has an approximate s-sparse representation in Φ, i.e. there exists a set $S \subseteq \mathcal{I}$ of size $|S| \le s$ for which $f \approx f_S = \sum_{l \in S} c_l \phi_l$ for $c_l \in \mathbb{V}$.

II. Main Results

Case (i): known S. We consider a positive weight function $w : D \to (0, \infty)$ and construct the approximation $\hat{f}$ to f via a Hilbert-valued weighted least-squares fit

$$\tilde{f} \in \operatorname{argmin}\{\mathcal{L}((p(y_i))_i, (b_i)_i) : p \in P_{S;\mathbb{V}_h}\}, \tag{1}$$

where $P_{S;\mathbb{V}} = \{\sum_{l \in S} c_l \phi_l \ : \ c_l \in \mathbb{V}\} \subset L^2_\rho(D;\mathbb{V})$ and $\mathcal{L}((p(y_i))_i, (b_i)_i) = \frac{1}{m} \sum_{i=1}^m w(y_i) \|b_i - p(y_i)\|^2_{\mathbb{V}}$. Our main result, stated informally for succinctness, is the following:

Theorem 2.1 (Optimal sampling; known S): There exists a choice of measures μ_i and weight function w such that f is recovered accurately and stably via (1) (with high probability), subject to the near-optimal sample complexity bound $m \gtrsim s \cdot \log(s)$. This bound is generally not achieved by MC sampling. Sample complexity bounds for MC sampling can be arbitrarily bad, depending on Φ and S.

This result extends previous work [1]–[6] to the Hilbert-valued setting. We observe that in practice the measures μ_i can be chosen as discrete measures, thus making it straightforward to draw samples from them. Note that by 'accurately' and 'stably' we mean f is recovered up to an error depending on $f - f_S$, the noise terms n_i and $f - \mathcal{P}_h(f)$, where $\mathcal{P}_h(f)$ is the orthogonal projection onto $\mathbb{V}_h$. This latter term accounts for the discretization of the space $\mathbb{V}$.

Case (ii): unknown S. We now further assume that $|\Phi| = n$ is finite and linearly independent, and consider the ℓ^1-minimization problem

$$\tilde{f} \in \operatorname{argmin}\{\lambda \|c\|_{\ell^1(\mathbb{V}^n)} + \sqrt{\mathcal{L}((p(y_i))_i, (b_i)_i)}\}, \tag{2}$$

where the minimization is taken over $p = \sum_{\iota \in \mathcal{I}} c_\iota \phi_\iota \in P_{\mathcal{I};\mathbb{V}_h}$. This is a Hilbert-valued version of the square-root LASSO problem [7]–[9]; the latter being particularly well suited to practical function approximation scenarios when the noise level is unknown [9]. Extending a number of previous works [9]–[15], our main result is:

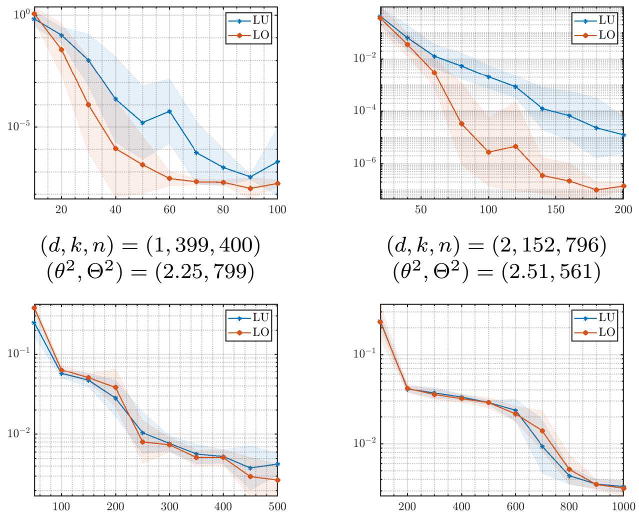
Theorem 2.2 (Towards optimal sampling; unknown S): There exists a choice of (discrete) measures μ_i and weight function w such that f is recovered accurately and stably via (2) (with high probability), subject to the sample complexity bound

$$m \gtrsim (b/a) \cdot (\theta^2/a) \cdot s \cdot \log(n) \cdot \log^2((b/a)(\theta^2/a)s),$$

where $a, b > 0$ are the Riesz basis bounds for Φ and $\theta^2 := \int_D \max_{\iota \in \mathcal{I}} |\phi_\iota(y)|^2 \, \mathrm{d}\rho(y)$. Conversely, the corresponding sample complexity bound for MC sampling involves the larger factor $\Theta^2 = \max_{\iota \in \mathcal{I}} \|\phi_\iota\|^2_{L^\infty_\rho(D)}$. Furthermore, there are choices of Φ for which $\theta = 1$ and Θ is arbitrarily large.

III. Conclusion

This work strives to understand optimal sampling for function approximation in general dictionaries; in particular, the extent to which one can improve standard MC sampling. It leads naturally to several new techniques, including a novel approach for function approximation on irregular domains. See Figs. 1–2 for numerical experiments. We remark that this approach can be significantly generalized, both in terms of the sampling and the low-dimensional structure. One can replace pointwise evaluations by sampling according to random linear operators, with potentially different and infinite-dimensional codomains. Further, one can replace the sparsity model by a structured sparsity model, for weighted [16], [17], lower set [14], [18] or joint sparsity. Extensions of Theorems 2.1 and 2.2 can be established for this substantially more general problem, leading to improved or optimal sampling strategies for other function approximation problems, such as dense-in-time, sparse-in-space sampling, gradient-augmented sampling [19]–[21] and numerous others.



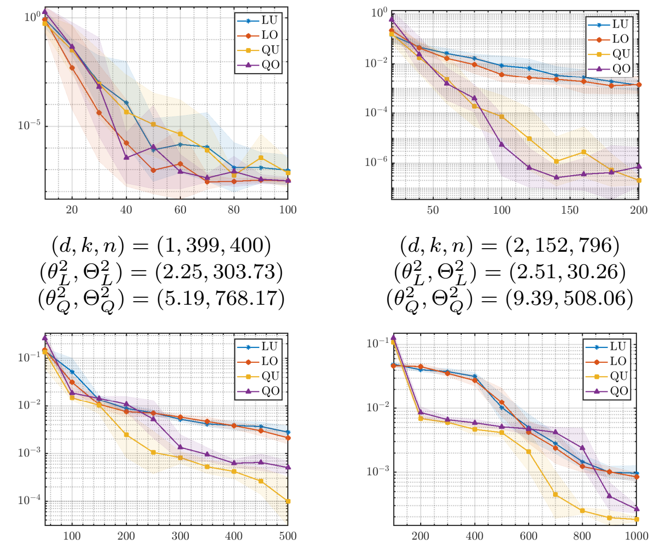
$(d, k, n) = (1, 399, 400)$
$(\theta^2, \Theta^2) = (2.25, 799)$

$(d, k, n) = (2, 152, 796)$
$(\theta^2, \Theta^2) = (2.51, 561)$

$(d, k, n) = (8, 22, 1843)$
$(\theta^2, \Theta^2) = (1.99, 81)$

$(d, k, n) = (16, 14, 4385)$
$(\theta^2, \Theta^2) = (1.74, 45)$

Fig. 1. Demonstrating the benefits of the sampling measures of Theorem 2.2. In this example, the dictionary is an orthonormal Legendre polynomial basis over $L^2_\rho(D)$, where $D = [-1, 1]^d$, ρ is the uniform measure and $\mathcal{I} = \mathcal{I}^{\mathrm{HC}}_{k-1}$ is the hyperbolic cross index set. The figures show the approximation error versus m for approximating the function $f(y) = \exp(-\sum_{k=1}^d y_k/2d)$ for different values of d over 10 trials, via MC sampling ('LU') and the sampling measures defined in Theorem 2.2 ('LO'). The values of the constants θ^2 and Θ^2 are also displayed. It is notable that the biggest improvement arises in lower dimensions, both theoretically (via the relative sizes of Θ and θ) and numerically (via the approximation error).



$(d, k, n) = (1, 399, 400)$
$(\theta_L^2, \Theta_L^2) = (2.25, 303.73)$
$(\theta_Q^2, \Theta_Q^2) = (5.19, 768.17)$

$(d, k, n) = (2, 152, 796)$
$(\theta_L^2, \Theta_L^2) = (2.51, 30.26)$
$(\theta_Q^2, \Theta_Q^2) = (9.39, 508.06)$

$(d, k, n) = (8, 22, 1843)$
$(\theta_L^2, \Theta_L^2) = (1.99, 5.77)$
$(\theta_Q^2, \Theta_Q^2) = (66.99, 3659)$

$(d, k, n) = (16, 14, 4385)$
$(\theta_L^2, \Theta_L^2) = (1.74, 2.94)$
$(\theta_Q^2, \Theta_Q^2) = (132, 11406)$

Fig. 2. A new procedure for polynomial approximation on irregular domains $D \subset [-1, 1]^d$, where $D = \{y : 1/4 \le \|y\|_2 \le 1\}$. In this example, we first consider the dictionary Φ formed by the restriction of the orthonormal Legendre basis from Figure 1 to D. We then consider the 'LU' and 'LO' strategies over D. Both lead to relatively poor approximations, since the dictionary Φ is poorly conditioned. As an alternative, we orthogonalize this basis over the support of the measures via QR factorization, then consider the two sampling strategies for this new basis (termed 'QU' and 'QO' respectively). Orthogonalization may destroy sparsity depending on how the original basis is ordered. To retain approximate sparsity, we order the basis according to increasing total order. On the other hand, in high dimensions, MC sampling actually outperforms the strategy of Theorem 2.2. This suggests further investigations are needed to obtain sampling strategies that consistently outperform MC sampling. The figures show the approximation error versus m for approximating the function $f(y) = \exp(-\sum_{k=1}^d y_k/2d)$ for different values of d over 10 trials.

References

[1] J. Hampton and A. Doostan, "Coherence motivated sampling and convergence analysis of least squares polynomial chaos regression," *Comput. Methods Appl. Mech. Engrg.*, vol. 290, pp. 73–97, 2015.

[2] A. Cohen and G. Migliorati, "Optimal weighted least-squares methods," *SMAI J. Comput. Math.*, vol. 3, pp. 181–203, 2017.

[3] B. Adcock and J. M. Cardenas, "Near-optimal sampling strategies for multivariate function approximation on general domains," *SIAM J. Math. Data Sci.*, vol. 2, no. 3, pp. 607–630, 2020.

[4] G. Migliorati, "Adaptive approximation by optimal weighted least squares methods," *SIAM J. Numer. Anal.*, vol. 5, no. 57, pp. 2217–2245, 2019.

[5] A. Cohen and M. Dolbeault, "Optimal sampling and christoffel functions on general domains," *arXiv:2010.1104*, 2020.

[6] B. Arras, M. Bachmayr, and A. Cohen, "Sequential sampling for optimal weighted least squares approximations in hierarchical spaces," *SIAM J. Math. Data Sci.*, vol. 1, no. 1, pp. 189–207, 2019.

[7] A. Belloni, V. Chernozhukov, and L. Wang, "Pivotal estimation via square-root Lasso in nonparametric regression," *Ann. Statist.*, vol. 42, no. 2, pp. 757–788, 2014.

[8] ——, "Square-root lasso: pivotal recovery of sparse signals via conic programming," *Biometrika*, vol. 98, no. 4, pp. 791–806, 2011.

[9] B. Adcock, A. Bao, and S. Brugiapaglia, "Correcting for unknown errors in sparse high-dimensional function approximation," *Numer. Math.*, vol. 142, no. 3, pp. 667–711, 2019.

[10] J. Hampton and A. Doostan, "Compressive sampling of polynomial chaos expansions: convergence analysis and sampling strategies," *J. Comput. Phys.*, vol. 280, pp. 363–386, 2015.

[11] H. Rauhut and R. Ward, "Sparse Legendre expansions via ℓ_1-minimization," *J. Approx. Theory*, vol. 164, no. 5, pp. 517–533, 2012.

[12] N. Dexter, H. Tran, and C. Webster, "A mixed ℓ_1 regularization approach for sparse simultaneous approximation of parameterized PDEs," *ESAIM Math. Model. Numer. Anal.*, vol. 53, pp. 2025–2045, 2019.

[13] J. D. Jakeman, A. Narayan, and T. Zhou, "A generalized sampling and preconditioning scheme for sparse approximation of polynomial chaos expansions," *SIAM J. Sci. Comput.*, vol. 39, no. 3, pp. A1114–A1144, 2017.

[14] B. Adcock, S. Brugiapaglia, and C. G. Webster, "Compressed sensing approaches for polynomial approximation of high-dimensional functions," in *Compressed Sensing and its Applications: Second International MATHEON Conference 2015*, ser. Applied and Numerical Harmonic Analysis, H. Boche, G. Caire, R. Calderbank, M. März, G. Kutyniok, and R. Mathar, Eds. Cham: Birkhäuser, 2017, pp. 93–124.

[15] L. Yan, L. Guo, and D. Xiu, "Stochastic collocation algorithms using ℓ_1-minimization," *Int. J. Uncertain. Quantif.*, vol. 2, no. 3, pp. 279–293, 2012.

[16] H. Rauhut and R. Ward, "Interpolation via weighted ℓ_1 minimization," *Appl. Comput. Harmon. Anal.*, vol. 40, no. 2, pp. 321–351, 2016.

[17] J. Peng, J. Hampton, and A. Doostan, "A weighted ℓ_1-minimization approach for sparse polynomial chaos expansions," *J. Comput. Phys.*, vol. 267, pp. 92–111, 2014.

[18] A. Chkifa, N. Dexter, H. Tran, and C. G. Webster, "Polynomial approximation via compressed sensing of high-dimensional functions on lower sets," *Math. Comp.*, vol. 87, no. 311, pp. 1415–1450, 2018.

[19] B. Adcock and Y. Sui, "Compressive Hermite interpolation: sparse, high-dimensional approximation from gradient-augmented measurements," *Constr. Approx.*, vol. 50, pp. 167–207, 2019.

[20] J. Peng, J. Hampton, and A. Doostan, "On polynomial chaos expansion via gradient-enhanced l_1-minimization," *J. Comput. Phys.*, vol. 310, pp. 440–458, 2016.

[21] L. Guo, A. Narayan, and T. Zhou, "A gradient enhanced ℓ_1-minimization for sparse approximation of polynomial chaos expansions," *J. Comput. Phys.*, vol. 367, pp. 49–64, 2018.