Adaptive Partitioning Schemes for Black-Box Optimization

Raja Sunkara RS5CQ@MST.EDU
Ardhendu Tripathy ARDHENDUTR@GMAIL.COM
Missouri University of Science & Technology

Abstract

Applications such as engineering design and hyperparameter tuning often require us to optimize a black-box function, i.e., a system whose inner processing is not analytically known and whose gradients are not available. Practitioners often have a fixed budget for the number of function evaluations and the performance of an optimization algorithm is measured by its simple regret. In this paper, we study the class of "Optimistic Optimization" algorithms for black-box optimization that use a partitioning scheme for the domain. We develop algorithms that learn a good partitioning scheme and use flexible surrogate models such as neural networks in the optimization procedure. For multi-index functions on an m-dimensional subspace within d dimensions, our algorithm attains $\tilde{O}(n^{-\beta/d})$ regret, where $\beta = 1 + \frac{d-m}{2m-1}$, as opposed to $\tilde{O}(n^{-1/d})$ for SequOOL, a state-of-the-art optimistic optimization algorithm. In numerical experiments on benchmark functions, our algorithm converged using 21% to 36% fewer evaluations compared to SequOOL. Our approach improves the quality of activation aware quantization of the OPT-1.3B large language model.

1. Introduction and motivation

Optimization of black-box functions is often carried out by a class of algorithms called "Optimistic Optimization" algorithms [18]. These algorithms are preferred due to their mild assumptions; however, they require a partition scheme to be provided for the search space. The quality of a partitioning scheme depends on the function being optimized [7]. If information on the function is lacking, then a default partitioning scheme, i.e., axis-aligned rectangles, is used [1, 11]. But this default choice might not be a good choice for the function. Additionally, an axis-aligned rectangle scheme limits the application to low-dimensional search spaces (as the number of rectangles grows as d^h, where d is the dimension and h is the height of the partition tree), or it may fail to exploit low-dimensional structures in high-dimensional spaces. Thus, there is a need to develop partitioning schemes that can adapt to the low dimensional structure that may be present in a black-box function.

One of the first global optimization algorithms with provable convergence was obtained for the class of Lipschitz functions. The DiRect algorithm [10, 11] is a well-known algorithm that can optimize Lipschitz functions without knowing the Lipschitz constant. It partitions the domain $\mathcal{X}$ into axis-aligned rectangles and refines those rectangles which could potentially be the maximum. Because of its partitioning scheme, it is typically limited to low-dimensional domains. A different class of functions studied in the Bayesian Optimization literature is that of the Gaussian Process (GP) prior for the unknown function f. Combining the observed samples with the prior mean and covariance kernel, a posterior distribution for f is obtained and used to guide the sampling strategy, see e.g. GP-UCB [25]. Since its sampling strategy required maximizing the Upper Confidence Bound obtained from the posterior, its could feasibly be applied only on low-dimensional $\mathcal{X}$. Later works

© R. Sunkara & A. Tripathy.

[23, 24] incorporated domain partitioning ideas to reduce the computational cost. However, these methods require the kernel associated with f as an input.

Unlike the methods described above, which assume a global characteristic for f, the "Optimistic Optimization" class of algorithms (HOO [2], SOO [17], SequOOL [1]) just assume a local smoothness condition around the global maximizer. These algorithms require as input a hierarchical partitioning of $\mathcal{X}$ that is well-behaved with respect to a semi-metric on $\mathcal{X}$. Although the semi-metric is not needed to be known, the performance of these algorithms can be heavily influenced by the choice of the partitioning scheme. Absent any additional information, the default partitioning is the axis-aligned rectangles from DiRect, leading to the shortcomings described in the beginning. We propose to develop an algorithm that adaptively builds a partitioning scheme as new samples are collected.

Main contributions When the function is a low-dimensional multi-index function we theoretically prove improved regret bounds shown in Table 1. Empirically, we demonstrate the improvement in optimization error for several benchmark functions including Rastrigin (multi-modal), Branin (multiple minima), and Sharp Ridge (non-differentiable). We pose Large Language Model (LLM) quantization as a high-dimensional black-box optimization problem and obtain improved results.

	SOO	SequOOL	Our Method
$\eta = 0$	$\rho^{\sqrt{n}}$	$\rho^{\tilde{\Omega}(n)}$	$\rho^{\beta\tilde{\Omega}(n)}$
$\eta > 0$	$\tilde{O}(n^{-1/\eta})$	$\tilde{O}(n^{-1/\eta})$	$\tilde{O}(n^{-\beta/\eta})$

Table 1: Regret bounds on a budget of n evaluations for a m-dimensional multi-index function in d dimensions. $\beta = 1 + \frac{d-m}{2m-1}$ and ρ, η are parameters for the default partitioning scheme.

Related works Here we summarize some methods we have compared in our experiments. Perhaps the closest related work is Random Embedding Simultaneous Optimistic Optimization (RESOO) [22], which scales SOO to high-dimensional optimization problems by applying SOO in a random low-dimensional search space. The simple regret of RESOO depends only on the effective dimension of the problem, rather than the full dimension of the solution space. REMBO (Random EMbedding Bayesian Optimization) ([29]) uses a random projection matrix to create a lower-dimensional embedding for high-dimensional optimization problems. It then applies Bayesian optimization on this low-dimensional space, allowing it to efficiently search for optima in the reduced space. HeSBO ([19]) uses hashing-enhanced embeding subspaces. ALEBO (Adaptive Linear Embedding Bayesian Optimization) ([12]) builds upon and improves the original REMBO by proposing a new linear-embedding method. However, these algorithms requires an lower-bound to the low-dimensional subspace dimension, which is difficult to obtain in real-world problems. Additionally, the Bayesian Optimization algorithms can be computationally expensive for large budgets.

Latent Action Monte Carlo Tree Search (LA-MCTS) [28] recursively partitions the high-dimensional search space into regions with high/low function values using nonlinear decision boundaries. Their boundaries are adaptive to the function $f(\mathbf{x})$. It serves as a meta-algorithm by using existing black-box optimizers (e.g., BO, TuRBO [4]) as its local model.

Evolutionary algorithms such as CMA-ES [8] and simulated annealing [30] are other popular approaches for black-box optimization. CMA-ES technique perform well at finding the best solutions in high-dimensional optimization problems. However, a downside of these methods is that they do not come with convergence guarantees ([14], [20]).

2. Problem formulation and adaptive partitioning schemes

We consider the problem of optimizing a function $f\colon \mathcal{X} \mapsto \mathbb{R}$ using only its evaluations at appropriately chosen points in its domain $\mathcal{X}$, which is assumed to be a closed and compact set. Given a budget of n evaluations, at each $t \in \{1, 2, \ldots, n\}$ the algorithm queries the point $\mathbf{x}_t \in \mathcal{X}$ and observes a real number $y_t = f(\mathbf{x}_t)$. After exhausting its budget, the algorithm returns the estimated maximizer $\hat{\mathbf{x}}(n)$. The optimization error is quantified by the simple regret r_n, defined as

$$r_n \triangleq f^* - f(\hat{\mathbf{x}}(n)), \text{ where } f^* \triangleq \sup_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}), \text{ and } \mathbf{x}^* \in \mathcal{X} \text{ such that } f^* = f(\mathbf{x}^*). \tag{1}$$

Our focus is on the class of *optimistic optimization* algorithms [1, 2, 7, 17, 18, 27]. These algorithms require as input a hierarchical partitioning of the domain $\mathcal{X}$ for their search procedure.

Definition 1 *Partitioning scheme [1]. Let $\mathcal{P}$ denote a tree representation of the domain $\mathcal{X}$. All the cells at depth h form a partition of $\mathcal{X}$ and are denoted as $\mathcal{P}_h$. We index the cells at depth h with an additional index i, i.e., $\mathcal{P}_{h,i}$ is a cell in the tree at depth h. We use $\mathcal{P}_h^*$ to denote a cell at depth h containing a maximizer $\mathbf{x}^*$ of f. We also use $\mathbf{x}_{h,i}$ to denote a representative location within the $\mathcal{P}_{h,i}$ cell.*

A common choice of $\mathcal{X}$ is obtained when we have interval constraints on each of its components. In this case, without loss of generality, we can consider $\mathcal{X} = [-1, 1]^d$. And a default choice for the partitioning scheme that is often used in practice is the axis-aligned trisection scheme [11].

Definition 2 *The default partitioning scheme constructs a hierarchical partitioning $\mathcal{P} = \{\mathcal{P}_{h,i}\}_{h,i}$ of $\mathcal{X} = [-1, 1]^d$ using an axis-aligned trisection method in a round-robin manner. At depth $h = 0$, there is a single cell $\mathcal{P}_{0,1} = \mathcal{X}$. Each cell $\mathcal{P}_{h,i}$ at depth h is split into three children cells $\{\mathcal{P}_{h+1,j}\}_j$ at depth $h + 1$. The trisection occurs along the $(h \bmod d) + 1$ axis, i.e., the new cells are created by introducing $(d - 1)$-dimensional hyperplanes orthogonal to the chosen axis. The representative $\mathbf{x}_{h,i}$ is the midpoint of the cell $\mathcal{P}_{h,i}$ and $\{\mathcal{P}_{h,i}\}_{1\le i\le 3^h}$ partitions $\mathcal{X}$ at each depth h.*

We also consider partitioning schemes that are not aligned with the standard canonical basis. We can define such a rotated low-dimensional partitioning scheme using a matrix of orthonormal rows.

Definition 3 *Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times d}$ such that $\mathbf{A}\mathbf{A}^\top = \mathbf{I}_m$ and a scalar $\alpha > 0$, we establish a partitioning scheme denoted as $\mathcal{A}$. Let the default partitioning scheme (Definition 2) on $[-\alpha, \alpha]^m$ be denoted as $\mathcal{T}$. For any depth h and index i, $\mathcal{A}_{h,i} \triangleq \{\mathbf{A}^\top \mathbf{t} : \mathbf{t} \in \mathcal{T}_{h,i}\}$ is a cell in the partitioning on the m-dimensional projection of $\mathcal{X}$ onto the subspace spanned by rows of $\mathbf{A}$.*

Since the projection of $\mathcal{X}$ onto $\mathbf{A}$ can result in points outside $\mathcal{X}$, the value of α is chosen to ensure that the projection is covered by the $\mathcal{A}$ partitioning scheme. We characterize the benefit of using the partitioning scheme $\mathcal{A}$ for the class of multi-index functions [5], i.e., if there is a matrix $\mathbf{A} \in \mathbb{R}^{m\times d}$ with orthonormal rows and a Lipschitz function g such that

$$f(\mathbf{x}) = g(\mathbf{A}\mathbf{x}), \tag{2}$$

then using the partitioning scheme $\mathcal{A}$ can decrease r_n at a faster rate with n. Formally, we use the *near-optimality dimension* from [1] which characterizes the difficulty of optimizing a black-box function using a partitioning scheme (see Definition 13). Omitted details/proofs are in the Appendix.

The following example demonstrates that the partitioning scheme $\mathcal{A}$ has a lower near-optimality dimension than the default partitioning scheme for a simple function with $m = 1, d = 2$.

Example 4 *Consider the function* $f(x_1, x_2) = g(\mathbf{A}\mathbf{x}) = 1-|x_1|$ *with* $\mathbf{A} = [1, 0]$ *and* $g(z) = 1-|z|$. *Let* $\eta_{\mathcal{P}}, \eta_{\mathcal{A}}$ *be the near-optimality dimensions for the partitioning schemes* $\mathcal{P}, \mathcal{A}$ *from Definitions 2 and 3. Then we have that* $\eta_{\mathcal{P}} = 1$ *and* $\eta_{\mathcal{A}} = 0$.

In addition to identifying the important subspace spanned by m orthonormal directions, an adaptive partitioning scheme can choose which direction to split at each depth.

Definition 5 *Direction selection strategy. For a partitioning* $\mathcal{A}$ *in Definition 3, a direction selection strategy* $\tau_h : \mathcal{H} \to [1 : m]$ *defined for each height* h *takes as input the history* $\mathcal{H}$ *of all past function evaluations till depth* $h - 1$ *and outputs the index of the direction to be split at depth* h.

In the following example, a direction selection strategy that splits the x_1 axis twice as often as the x_2 axis leads to a better regret than the default partitioning scheme $\mathcal{P}$ which splits both the axes in equal proportion.

Example 6 *The near-optimality dimension of the default partitioning scheme for the function* $f(x_1, x_2) = 1 - |x_1| - x_2^2$ *is* $\eta_{\mathcal{P}} = 0.5$. *On the other hand, consider the partitioning scheme* $\mathcal{A}$ *from Definition 5 with* $\mathbf{A} = I_2, \alpha = 1$ *and direction selection strategy* $\tau_h = 1$ *if* $h \mod 3 \neq 0$ *and* $\tau_h = 2$ *otherwise. Its near-optimality dimension* $\eta_{\mathcal{A}} = 0$.

3. Proposed algorithms for black-box optimization

We develop two kinds of algorithms: 1. a two-stage algorithm where the first stage learns an adaptive partitioning scheme and the second stage uses it for optimization, and 2. an interleaved algorithm where learning and optimization happen iteratively.

Two-stage algorithm. In the first stage, we use a learning algorithm to obtain $\hat{\mathbf{A}}$, i.e., the directions used to define the adaptive partitioning scheme $\mathcal{A}$. If f is a multi-index function satisfying (2), the quality of this estimate is measured by the subspace distance between $\hat{\mathbf{A}}$ and the true $\mathbf{A}$. In that case, the value of $\hat{\alpha}$ is chosen in Lemma 23 such that $f(\hat{\mathbf{A}}^\top \mathbf{t}) = f(\mathbf{x}^\star)$ for some $\mathbf{t} \in [-\hat{\alpha}, \hat{\alpha}]^m$ and the optimization can converge to the maximizer in the low-dimensional subspace.

Algorithm 1: Obtaining directions for an adaptive partitioning scheme

Input: T, m, oracle for f which is a multi-index function defined using $\mathbf{A}$ (see (2)).
Sample f at T points chosen as $x^{(i)} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and fit a single hidden-layer neural network $\hat{y}$;
$\hat{\mathbf{A}} \leftarrow$ top m right singular vectors of the hidden layer weight matrix;
$u \leftarrow$ Upper bound to $\operatorname{dist}(\mathbf{A}, \hat{\mathbf{A}})$ obtained in lemma 31 or theorem 32;
return $\hat{\mathbf{A}}$ *and* $\hat{\alpha} = \sqrt{dm/(1-u^2)}$ used to specify the partitioning scheme $\mathcal{A}$ in Definition 3.

Algorithm 1 learns $\hat{\mathbf{A}}$ by fitting a single hidden-layer neural network to the function evaluations at random locations in its domain. A single hidden-layer neural network can model the fact that only a subspace of the domain explains all the variation in the function values (see Proposition 11). In practice, we can choose the value of m to explain a desired percentage (such as 95%) of the total variation in the SVD step calculating $\hat{\mathbf{A}}$. Another approach we consider is learning $\hat{\mathbf{A}}$ from Fornasier et al. [5, Algorithm 2] which estimates the gradient of the function using finite differences. The second stage applies SequOOL to the partitioning scheme $\mathcal{A}$ returned by Algorithm 1.

Interleaved learning and optimization. Instead of separating the learning and optimization in two distinct stages, an interleaved algorithm updates the neural network fit at regular intervals. The updated approximant is used to specify the direction selection strategy in Definition 5.

Algorithm 2: SequOOL on an adaptive partitioning scheme with a direction selection strategy

Input: Total number of openings n, number of samples T for updating $\hat{f}$, integer c stating how often $\hat{f}$ is updated, number of dimensions m, oracle for f, direction selection strategy τ_h.
Setup a partitioning scheme $\hat{\mathcal{A}}$ with $\hat{\mathbf{A}}$ and $\hat{\alpha}$ obtained using [5, Algorithm 2];
Obtain $\hat{f}$ by fitting a neural network on available samples, set $h_{\max} \leftarrow \lfloor n^2 / \left(n\overline{\log} n + \frac{Tn}{3c}\right) \rfloor$;
for $h \leftarrow 1$ **to** $h_{\max}$ **do**
  $\mathcal{T}_h^* \leftarrow$ the cell with the largest function sample value at height h;
  if $h \mod c = 0$ **then**
    Obtain evaluations of f at T uniform random locations in $\mathcal{T}_h^*$ and update $\hat{f}$;
  Open $\lfloor h_{\max}/h \rfloor$ cells at depth h and trisect them along the direction returned by $\tau_h(\hat{f})$;
return $\hat{\mathbf{x}}(n) \leftarrow \arg\max_{all\ h,i} f(\mathbf{x}_{h,i})$

Algorithm 2 uses the parameter τ_h that takes the updated approximation $\hat{f}$ as input. Our proposed method for τ_h is the lookahead direction selection strategy, which is detailed in Algorithm 3 in Appendix section 3.

Proposition 7 *Let $n, T, c, h_{\max}$ be the parameters defined in Algorithm 2 with $\overline{\log} n \triangleq \sum_{t=1}^{n} \frac{1}{t}$. Then the total number of function evaluations taken by the algorithm will not exceed $3n$.*

For the regret upper bound of our Algorithm 1, refer to the theoretical section in Appendix D.

4. Experiments

We evaluate our algorithms against state-of-the-art baselines from various optimization categories. These include Bayesian Optimization (REMBO [29], HesBO [19]), Evolutionary Algorithms (CMA-ES [9]), Dual Annealing [21], Optimistic Optimization (SOO, SequOOL, RESOO [22], DiRect), and Random Search. The optimization functions used in our experiments include Sphere, Rastrigin, Different Powers, Rosenbrock, Styblinski-Tang, Hartmann-6, Branin, Ellipsoid, Sharp Ridge, and the CUSTOM function defined as $1 + (x_1 - 1)^2 + \sum_{i=d-m+2}^{d} (x_i - 1)^4$.

In our experimental setup, we construct the multi-index function as $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^d$, $g : \mathbb{R}^m \to \mathbb{R}$ is a standard optimization test function, and $\mathbf{A} \in \mathbb{R}^{m \times d}$ is a randomly generated matrix satisfying $\mathbf{A}\mathbf{A}^\top = \mathbf{I}_m$. The Branin and Hartmann-6 functions are defined in 2 and 6 dimensions respectively; thus, we choose $m = 2$ for Branin and $m = 6$ for Hartmann-6 when constructing the multi-index function. To further evaluate the efficacy of our algorithm, we conducted benchmarks on low-dimensional multi-index functions with $d = 5$ and $m = 2$. Additional experimental results are in Appendix G.2.

Our algorithm demonstrates superior performance on the Rastrigin, Sphere and Styblinski-Tang functions, achieving zero regret with fewer samples compared to other methods. On the different powers function, we perform comparable to the other methods, however, on the $(x_1 - 1)^2 + (x_{72} - 1)^4$, Ellipsoid and Sharp Ridge function, our algorithm perform slightly worse than the RESOO. This may

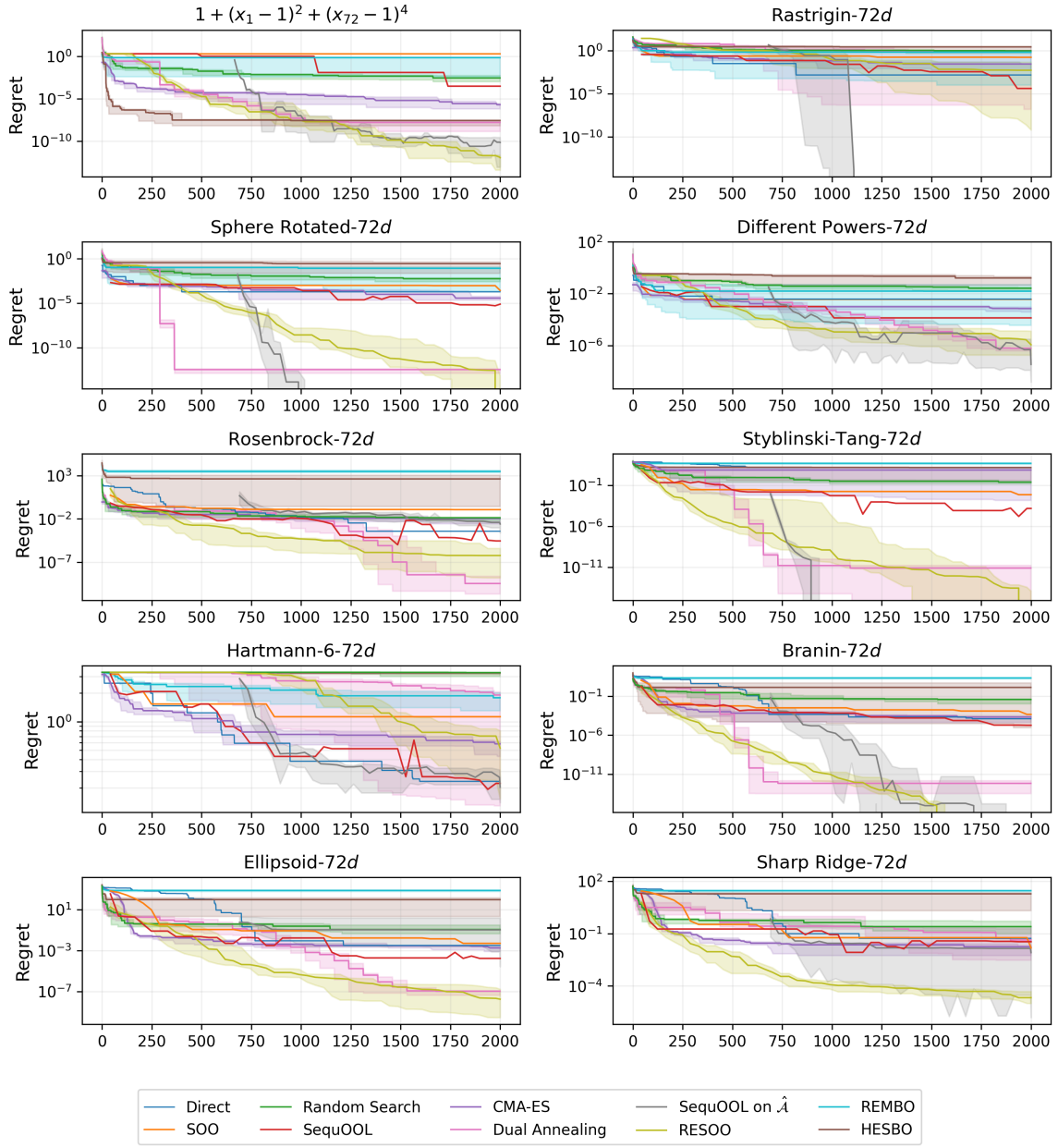

Figure 1: Regret Plots: Algorithm 1 (SequOOL on $\hat{\mathcal{A}}$) uses 650 additional samples to learn the subspace. Regret is plotted for 100 equally spaced budget values between 1 and 2000. For the randomized algorithms, we took 10 trials and plotted the median curve (thick line) and 0 and 95 percentile curves.

be attributed to the use of several (650) samples for the first stage in a 2000 budgeted experiment. Experiment on the LLM Quantization are in Appendix G.3.

References

[1] Peter L. Bartlett, Victor Gabillon, and Michal Valko. A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption. In Aurélien Garivier and Satyen Kale, editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 184–206. PMLR, 22–24 Mar 2019. URL https://proceedings.mlr.press/v98/bartlett19a.html.

[2] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(5), 2011.

[3] Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

[4] David Eriksson, Michael Pearce, Jacob Gardner, Ryan D Turner, and Matthias Poloczek. Scalable global optimization via local bayesian optimization. *Advances in neural information processing systems*, 32, 2019.

[5] Massimo Fornasier, Karin Schnass, and Jan Vybiral. Learning functions of few arbitrary linear parameters in high dimensions. *Foundations of Computational Mathematics*, 12:229–262, 2012.

[6] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[7] Jean-Bastien Grill, Michal Valko, and Rémi Munos. Black-box optimization of noisy functions with unknown smoothness. *Advances in Neural Information Processing Systems*, 28, 2015.

[8] Nikolaus Hansen. The cma evolution strategy: A tutorial, 2023. URL https://arxiv.org/abs/1604.00772.

[9] Nikolaus Hansen, Sibylle D Müller, and Petros Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es). *Evolutionary computation*, 11(1):1–18, 2003.

[10] Donald R Jones and Joaquim RRA Martins. The direct algorithm: 25 years later. *Journal of global optimization*, 79(3):521–566, 2021.

[11] Donald R Jones, Cary D Perttunen, and Bruce E Stuckman. Lipschitzian optimization without the lipschitz constant. *Journal of optimization Theory and Applications*, 79:157–181, 1993.

[12] Ben Letham, Roberto Calandra, Akshara Rai, and Eytan Bakshy. Re-examining linear embeddings for high-dimensional bayesian optimization. *Advances in neural information processing systems*, 33:1546–1558, 2020.

[13] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6:87–100, 2024.

[14] Ilya Loshchilov and Frank Hutter. Cma-es for hyperparameter optimization of deep neural networks. *arXiv preprint arXiv:1604.07269*, 2016.

[15] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[16] Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6taykzqcPD.

[17] Rémi Munos. Optimistic optimization of a deterministic function without the knowledge of its smoothness. *Advances in neural information processing systems*, 24, 2011.

[18] Rémi Munos. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7(1):1–129, 2014. ISSN 1935-8237. doi: 10.1561/2200000038. URL http://dx.doi.org/10.1561/2200000038.

[19] Amin Nayebi, Alexander Munteanu, and Matthias Poloczek. A framework for bayesian optimization in embedded subspaces. In *International Conference on Machine Learning*, pages 4752–4761. PMLR, 2019.

[20] Masahiro Nomura, Shuhei Watanabe, Youhei Akimoto, Yoshihiko Ozaki, and Masaki Onishi. Warm starting cma-es for hyperparameter optimization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 9188–9196, 2021.

[21] Martin Pincus. A monte carlo method for the approximate solution of certain types of constrained optimization problems. *Operations research*, 18(6):1225–1228, 1970.

[22] Hong Qian and Yang Yu. Scaling simultaneous optimistic optimization for high-dimensional non-convex functions with low effective dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[23] Sudeep Salgia, Sattar Vakili, and Qing Zhao. A domain-shrinking based bayesian optimization algorithm with order-optimal regret performance. *Advances in Neural Information Processing Systems*, 34:28836–28847, 2021.

[24] Shubhanshu Shekhar and Tara Javidi. Gaussian process bandits with adaptive discretization. *Electronic Journal of Statistics*, 12(2):3829–3874, 2018.

[25] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias W Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE transactions on information theory*, 58(5):3250–3265, 2012.

[26] Vinod Vaikuntanathan. Csc 2414 lattices in computer science. https://people.csail.mit.edu/vinodv/COURSES/CSC2414-F11/L1.pdf, 2011. [Online; accessed 19-July-2008].

[27] Michal Valko, Alexandra Carpentier, and Rémi Munos. Stochastic simultaneous optimistic optimization. In *International Conference on Machine Learning*, pages 19–27. PMLR, 2013.

[28] Linnan Wang, Rodrigo Fonseca, and Yuandong Tian. Learning search space partition for black-box optimization using monte carlo tree search. *Advances in Neural Information Processing Systems*, 33:19511–19522, 2020.

[29] Ziyu Wang, Frank Hutter, Masrour Zoghi, David Matheson, and Nando De Feitas. Bayesian optimization in a billion dimensions via random embeddings. *Journal of Artificial Intelligence Research*, 55:361–387, 2016.

[30] Yang Xiang, Sylvain Gubian, Brian Suomela, and Julia Hoeng. Generalized simulated annealing for global optimization: the gensa package. *R J.*, 5(1):13, 2013.

[31] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Appendix A. Notation

Let $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ denote the set of non-negative integers and $[d]$ represents the set $\{1, 2, \ldots, d\}$. For vectors and matrices, we use $(\cdot)^\top$ to denote the transpose. For a d-dimensional vector $\mathbf{x} = [x_1, \ldots, x_d]^\top$, we use $\|\mathbf{x}\|_p$ and $\|\mathbf{x}\|_\infty$ to denote its ℓ_p and ℓ_∞ norm respectively. For any matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, $\|\mathbf{X}\|_2$ and $\|\mathbf{X}\|_F$ represent its spectral and Frobenius norms. We use $\sigma_i(\mathbf{X})$ for the ith largest singular value, with $\sigma_{\max}(\mathbf{X})$ and $\sigma_{\min}(\mathbf{X})$ denoting the largest and smallest singular values. Suppose $\mathbf{A} \in \mathbb{R}^{m \times d}$, with $\mathbf{A}\mathbf{A}^\top = \mathbf{I}_m$, then for any vector $\mathbf{v} \in \mathbb{R}^d$, we denote its orthogonal projection onto the span of the rows of $\mathbf{A}$ as $\mathbf{v}_\| = \mathbf{A}^\top \mathbf{A}\mathbf{v}$, with the orthogonal component given by $\mathbf{v}_\perp = \mathbf{v} - \mathbf{v}_\|$. For a matrix $\mathbf{W} \in \mathbb{R}^{p \times d}$, we denote $\mathbf{W}_\|$ and $\mathbf{W}_\perp$ as the projections applied to each row, i.e., $\mathbf{W}_\| = \mathbf{W}\mathbf{A}^\top\mathbf{A}$ and $\mathbf{W}_\perp = \mathbf{W} - \mathbf{W}_\|$. For a given scalar $\kappa > 0$, we denote $\kappa\mathcal{X}$ as the set $\{\kappa x : x \in \mathcal{X}\}$. By a partitioning scheme $\mathcal{P}$ with κ, we mean the domain for the partitioning scheme is the set $\{\kappa x : x \in [-1, 1]^d\}$. By the default partitioning scheme $\mathcal{P}$, we mean partitioning scheme $\mathcal{P}$ with $\kappa = 1$. To describe side-length of a hyper-rectangle, we use the notation $[3^{-\lfloor \frac{h+m-i}{m} \rfloor}]_{i=1}^m$. This expressions represents a vector of m components, where each component is given by $3^{-\lfloor \frac{h+m-i}{m} \rfloor}$, with i ranging from 1 to m.

Appendix B. Omitted details for Section 2

Definition 8 *Let $\mathbf{c}_j \in \{-1, 1\}^d$ denote the 2^d corners, indexed by j, of the default axis-aligned $\mathcal{P}$ partitioning scheme. Given a matrix $\mathbf{A}$ with m orthonormal rows denoted as $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$, we define*

$$\boldsymbol{\alpha}_{\max} \triangleq [\max_{1 \le j \le 2^d} \mathbf{a}_1^\top \mathbf{c}_j, \max_{1 \le j \le 2^d} \mathbf{a}_2^\top \mathbf{c}_j, \cdots \max_{1 \le j \le 2^d} \mathbf{a}_m^\top \mathbf{c}_j]^\top$$

as the extent parameter of the $\mathcal{A}$ partitioning scheme. Additionally, we define the largest component of the extent parameter as $\alpha \triangleq \max_{1 \le i \le m} \max_{1 \le j \le 2^d} \mathbf{a}_i^\top \mathbf{c}_j$.

Proposition 9 *Suppose $f : \mathbb{R}^d \mapsto \mathbb{R}$ is a multi-index function of the form $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$ and we want to optimize it over $\mathcal{X} = [-1, 1]^d$. If $\mathbf{x}^* \in \mathcal{X}$ is an optimizer and α is chosen as per definition 8, then there exists a $\mathbf{z}^* \in \mathbb{R}^m$ such that $f(\mathbf{A}^\top \mathbf{z}^*) = f(\mathbf{x}^*)$ and $\|\mathbf{z}^*\|_\infty \le \alpha$.*

The proposition above implies that if we have access to true subspace matrix $\mathbf{A}$, we can compute α and restrict our optimization to the m dimensional space $\alpha\mathbb{H}_1^m$ and perform optimization on $\mathcal{A}$ partitioning scheme. This restriction guarantees that we can still recover the optimal function value f^* by optimizing over this lower-dimensional space. We characterize the benefit of using the partitioning scheme $\mathcal{A}$ for the class of multi-index functions, i.e., if f satisfies (2) then using the partitioning scheme $\mathcal{A}$ can decrease r_n at a faster rate with n. Formally, we use the *near-optimality dimension* from [1] which characterizes the difficulty of optimizing a black-box function using a partitioning scheme (see Definition 13).

Bartlett et al. [1] make a key assumption about the partitioning scheme that states that the suboptimality of any point in the $\mathcal{P}_h^*$ cell keeps improving as the depth h increases (see Assumption 12). The rate of this improvement is characterized by a parameter $\rho \in (0, 1)$. In the next example, the subspace defined by $\mathbf{A}$ is not aligned with the canonical axes. In this case, although the near-optimality dimension for both $\mathcal{P}$ and $\mathcal{A}$ is 0, we see that the $\rho_{\mathcal{A}}$ is smaller than the ρ of the default partitioning scheme, leading to a faster decrease in regret when using $\mathcal{A}$ instead of $\mathcal{P}$.

Example 10 *Consider the function* $f(x_1, x_2) = g(\mathbf{A}\mathbf{x}) = 1 - |x_1 + x_2|$ *with* $\mathbf{A} = [1, 1]$ *and* $g(k) = 1 - |k|$. *Let* $\eta_{\mathcal{P}}, \eta_{\mathcal{A}}$ *be the near-optimality dimensions for the partitioning schemes* $\mathcal{P}, \mathcal{A}$ *respectively. And* $(\nu, \rho), (\nu_{\mathcal{A}}, \rho_{\mathcal{A}})$ *be the parameters for the paritioning schemes* $\mathcal{A}, \mathcal{P}$ *respectively. Then,* $\rho = 1/\sqrt{3}$ *and* $\rho_{\mathcal{A}} = 1/3$ *and* $\eta_{\mathcal{P}} = \eta_{\mathcal{A}} = 0$.

Appendix C. Omitted details for Section 3

A single hidden layer neural network with p hidden neurons maps input $\mathbf{x} \in \mathbb{R}^d$ to the scalar

$$\hat{y}(\mathbf{x}, \mathbf{W}, \mathbf{a}, \mathbf{b}) = \sum_{i=1}^{p} a_i \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i), \tag{3}$$

where σ is the non-linear activation function, $\mathbf{W}$ is the hidden layer weight matrix consisting of p weight vectors denoted as $\mathbf{w}_i$, b_i is the scalar bias for the ith hidden neuron, and a_i are the components of the output layer weight vector. The following proposition shows that the class of single hidden layer neural networks can represent the important subspace of multi-index functions.

Proposition 11 *Consider a function* $f(\mathbf{x}) = \sum_{i=1}^{p} v_i \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i)$, *where* $\mathbf{x} \in \mathbb{R}^d$ *and* σ *is a non-linear function. Let* $\mathbf{x} = \mathbf{P}\mathbf{x} + (\mathbf{I} - \mathbf{P})\mathbf{x}$, *where* $\mathbf{P}$ *is the projection matrix that maps any* $x \in \mathbb{R}^d$ *to* $\mathrm{Span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p\}$. *Then* $f(\mathbf{x}) = \sum_{i=1}^{p} v_i \sigma(\mathbf{w}_i^\top \mathbf{P}\mathbf{x} + b_i)$. *And if* $\mathbf{x}, \mathbf{x'} \in \mathbb{R}^d$ *are such that* $(\mathbf{x} - \mathbf{x'}) \perp \mathrm{Span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p\}$, *then* $f(\mathbf{x}) = f(\mathbf{x'})$.

We use SGD with weight decay to fit the neural network on the function evaluations obtained at uniform random locations in $\mathcal{X}$. The top m right singular vectors of the learned weight matrix of the hidden layer is used as the estimated $\hat{\mathbf{A}}$ for obtaining the partitioning scheme $\hat{\mathcal{A}}$.

Algorithm 3 describes our lookahead direction selection strategy $\tau_h(\hat{f})$.

Algorithm 3: Implementing lookahead direction selection strategy $\tau_h(\hat{f})$

Input: Current partition tree $\mathcal{T}$, height h, estimated function $\hat{f}$
Result: An integer in $[1 : m]$ denoting $axis_to_split$
$\hat{\mathbf{x}}_h^* \leftarrow \arg\max_i f(\mathbf{x}_{h,i})$, representative of cell $\mathcal{T}_h$ with the largest function value at height h ;
$\mathcal{T}_h \leftarrow$ cell at height h in $\mathcal{T}$ whose representative is $\hat{\mathbf{x}}_h^*$, $axis_to_split \leftarrow 0$, $minimum \leftarrow \infty$;
for $i \leftarrow 1$ **to** m **do**
  $\mathcal{T}_{h+1} \leftarrow$ child cell of $\mathcal{T}_h$ after trisecting axis i and having the same representative $\hat{\mathbf{x}}_h^*$;
  temp $\leftarrow$ Compute minimum of $\hat{f}$ on the domain $\mathcal{T}_{h+1}$;
  if *temp* < *minimum* **then**
    $axis_to_split \leftarrow i$;
    $minimum \leftarrow temp$;
return $axis_to_split$;

Querying outside the domain and optimizing $\hat{f}$ In practice, with our optimization domain set as $\mathcal{X} = [-1, 1]^d$, we may encounter situations during low-dimensional subspace optimization where $\mathbf{t} \in [-\alpha, \alpha]^m$ results in $\hat{\mathbf{A}}^\top \mathbf{t} \notin \mathcal{X}$. To ensure that f can be evaluated in all cases, we employ a two-step approach. First, we attempt to solve the optimization problem: $\arg\min_{\mathbf{t}_c \in \mathbb{R}^{d-m}} \left\| \hat{\mathbf{A}}_c^\top \mathbf{t}_c \right\|_2$

subject to $\hat{\mathbf{A}}^\top \mathbf{t} + \hat{\mathbf{A}}_c^\top \mathbf{t}_c \in \mathcal{X}$, where $\hat{\mathbf{A}}_c$ consists of the remaining $d-m$ columns of $\hat{\mathbf{A}}$ that are not in $\hat{\mathbf{A}}$. If this optimization problem has no feasible solution, we then employ Euclidean projection onto $\mathcal{X}$: $\arg\min_{x\in\mathcal{X}} \left\|\mathbf{x} - \hat{\mathbf{A}}^\top \mathbf{t}\right\|_2$. This projection method is applied whenever $\hat{\mathbf{A}}^\top \mathbf{t} \notin \mathcal{X}$, ensuring that we always have a valid point within our optimization domain.In practice, we estimate the minimum of $\hat{f}$ on the domain $\mathcal{T}_{h+1}$ using random sampling or any of the other black-box optimization algorithms, since $\hat{f}$ is cheap to evaluate and gradients are also available.

Appendix D. Theoretical analysis

We use the assumption made by Grill et al. [7] and Bartlett et al. [1] which states that the suboptimality of any point in a cell containing the global maximizer strictly improves with increasing height.

Assumption 12 *[1] For any global optimum $x^\star$, there is a $\nu > 0$ and $\rho \in (0,1)$ such that for all $h \in \mathbb{N}_0$ and all $\mathbf{x} \in \mathcal{P}_h^*$, we have that $f(x) \geq f(x^\star) - \nu\rho^h$.*

Definition 13 *Near-optimality dimension from [1]. Consider a partitioning scheme $\mathcal{P}$ that satisfies Assumption 12 for some ν, ρ. For any $C > 1$, the near-optimality dimension $\eta_{\mathcal{P}}(\nu, \rho, C)$ of f with respect to the partitioning $\mathcal{P}$ is defined as $\eta_{\mathcal{P}}(\nu, \rho, C) \triangleq \inf\{\eta \in \mathbb{R}^+ : \forall h \geq 0, \mathcal{N}_{\mathcal{P}}(h) \leq C\rho^{-\eta h}\}$, where $\mathcal{N}_{\mathcal{P}}(h)$ is the number of cells $\mathcal{P}_{h,i}$ at depth h for which $\sup_{\mathbf{x}\in\mathcal{P}_{h,i}} f(\mathbf{x}) \geq f(\mathbf{x}^*) - 3\nu\rho^h$.*

Intuitively, a larger ρ implies that the function is only improving slowly near the maximizer, and a larger η implies that there are many near-optimal cells which must be ruled out to get the true maximizer. In both cases, we need a larger budget of evaluations to converge. We show that for the class of multi-index functions (2), the partitioning $\mathcal{A}$ has a lower η and a lower ρ compared to that of the default partitioning $\mathcal{P}$.

In this section, we establish relationships between three partitioning schemes: the default scheme $\mathcal{P}$, the scheme $\mathcal{A}$ based on the true subspace, and the scheme $\hat{\mathcal{A}}$ based on the estimated subspace. Our analysis proceeds in two main stages: We first relate the parameters of the default partitioning scheme $\mathcal{P}$ to those of the scheme $\mathcal{A}$. This includes comparing their SequOOL parameters (ν, ρ), characterizing their optimal cells, and bounding the number of near-optimal cells. We then extend this analysis to the estimated scheme $\hat{\mathcal{A}}$, relating its properties to those of $\mathcal{A}$. This involves quantifying the impact of using an estimated subspace and establishing relationships between the SequOOL parameters of $\hat{\mathcal{A}}$ and $\mathcal{A}$.

First, we start with relating the default partitioning scheme $\mathcal{P}$ with $\mathcal{A}$ partitioning scheme. To establish the relationship between the near-optimality dimensions of the two schemes, we first need to compare the parameters (ν, ρ) of `SequOOL` across both partitioning schemes. This requires a characterization of the cells $\mathcal{A}_h^*$and $\mathcal{P}_h^*$. The following proposition provides this characterization.

Proposition 14 *Let $\kappa > 0$ and $\boldsymbol{\alpha}_h^* \in \mathbb{R}^m$ be the representative of the $\mathcal{A}_h^*$ cell containing a global maxima of the function. Using fraction of two vectors to denote component-wise division,*

$$\mathcal{A}_h^* = \{\mathbf{A}^\top \boldsymbol{\alpha} : \boldsymbol{\alpha} \in \mathbb{R}^m, \left\|\frac{\boldsymbol{\alpha} - \boldsymbol{\alpha}_h^*}{\mathbf{s}}\right\|_\infty \leq \alpha \quad \textit{with} \quad \mathbf{s} = [3^{-\lfloor \frac{h+m-i}{m} \rfloor}]_{i=1}^m\}. \tag{4}$$

Similarly, if $\mathbf{x}_h^* \in \mathbb{R}^d$ is the representative of the $\mathcal{P}_h^*$ cell containing the same global maxima,

$$\mathcal{P}_h^* = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^d, \left\|\frac{\mathbf{x} - \mathbf{x}_h^*}{\mathbf{c}}\right\|_\infty \leq \kappa \quad \textit{with} \quad \mathbf{c} = [3^{-\lfloor \frac{h+d-i}{d} \rfloor}]_{i=1}^d\}. \tag{5}$$

Proposition 9 shows that we can use $\mathcal{A}$ partitioning scheme to perform optimization. We now relate the ν and ρ parameters (see Assumption 12) of the partitioning schemes $\mathcal{P}$ and $\mathcal{A}$.

To establish relationships between the parameters of the partitioning schemes $\mathcal{A}$ and the default scheme $\mathcal{P}$, we need to connect the sets $\mathcal{P}_h^*$ and $\mathcal{A}_h^*$. The following lemma provides this connection:

Lemma 15 *Suppose $\mathbf{x}^* \in \mathcal{P}_h^*$ is such that $\mathbf{x}^* = \mathbf{A}^\top \mathbf{A}\mathbf{x}^*$. If the domain for $\mathcal{P}$ is $\kappa\mathbb{H}_1^d$ with $\kappa \geq \sqrt{m}\alpha$, then $\forall i \in [1:m-1], \forall k \in \mathbb{N}_0$ we have that $\mathcal{A}_{km+i}^* \subseteq \mathcal{P}_{kd+i}^*$.*

Having established the relationship between the star cells of the partitioning schemes $\mathcal{A}$ and $\mathcal{P}$, we can now proceed to relate their respective parameters. The following two lemmas establish these relationships.

This lemma connects the parameters (ν, ρ) of the partitioning scheme $\mathcal{P}$ with $\kappa \geq \sqrt{m}\alpha$ to the parameters $(\nu_{\mathcal{A}}, \rho_{\mathcal{A}})$ of the scheme $\mathcal{A}$.

Lemma 16 *Let the parameters for the partitioning schemes $\mathcal{P}, \mathcal{A}$ be $(\nu, \rho), (\nu_{\mathcal{A}}, \rho_{\mathcal{A}})$ respectively. If Lemma 15 is applicable and $\mathcal{P}$ satisfies Assumption 12. Then we have that $\nu_{\mathcal{A}} = \nu\rho^{(1-\beta)(m-1)}, \rho_{\mathcal{A}} = \rho^\beta$ where $\beta = 1 + \frac{d-m}{2m-1}$.*

This lemma establishes the relationship between the parameters $(\nu_{\mathcal{A}}, \rho_{\mathcal{A}})$ of scheme $\mathcal{A}$ and (ν, ρ) of the default partitioning scheme $\mathcal{P}$.

Lemma 17 *The parameters $(\nu_{\mathcal{A}}, \rho_{\mathcal{A}}), (\nu, \rho)$ associated with partitioning schemes $\mathcal{A}$ and $\mathcal{P}$ with $\kappa = 1$. Let $l_f = f^* - \inf_{\mathbf{x}\in\kappa_1\mathbb{H}_1^d} f(\mathbf{x})$. Then*

$$\rho_{\mathcal{A}} = \rho^\beta, \nu_{\mathcal{A}} = \max\{\nu, l_f\}\rho^{(1-\beta)(m-1)-\tilde{h}_1}$$

where $\tilde{h}_1 = d\left\lceil \log_3 \kappa_1 \right\rceil, \beta = 1 + \frac{d-m}{2m-1}$ and $\kappa_1 = \sqrt{m}\alpha$.

The previous lemmas established relationships between the parameters of the partitioning schemes $\mathcal{A}$ and $\mathcal{P}$. They demonstrate that $\mathcal{A}$ is a valid partitioning scheme with a reduced sequOOL parameter $\rho_{\mathcal{A}}$ compared to the default partitioning scheme ρ.

Building on these results, we now turn our attention to comparing the number of near-optimal cells in each scheme.

It is helpful to use the notation of lattices to relate the number of near-optimal cells in two different partitioning schemes.

Definition 18 *[26] Given m linearly independent vectors $\mathbf{b}_1, \ldots, \mathbf{b}_m \in \mathbb{R}^m$, the lattice generated by them is defined as $L(\mathbf{b}_1, \ldots, \mathbf{b}_m) = \{\sum_{i=1}^m a_i\mathbf{b}_i \mid a_i \in \mathbb{Z}\}$. We call $\mathbf{b}_1, \ldots, \mathbf{b}_m$ a basis of the lattice. We denote lattices formed by the standard basis vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$ as Λ. Thus, $\Lambda = L(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m) = \mathbb{Z}^m$. The lattice Λ scaled by a scalar κ is the same as $L(\kappa\mathbf{e}_1, \kappa\mathbf{e}_2, \ldots, \kappa\mathbf{e}_n)$.*

For a d-dimensional vector $\mathbf{x} = [x_1, \ldots, x_d]^\top$, we use $\|\mathbf{x}\|_p$ and $\|\mathbf{x}\|_\infty$ to denote its ℓ_p and ℓ_∞ norm respectively.

Lemma 19 *Let $\kappa_1, \kappa_2 \in \mathbb{R}^+$ with $\kappa_1 > \kappa_2$, and let $B(\mathbf{x}, r) = \{\mathbf{y} \in \mathbb{R}^m : \|\mathbf{y} - \mathbf{x}\|_\infty \leq r\}$. Consider a lattice Λ as defined in Definition 18 , scaled by κ_2, and translated by a vector $\mathbf{t}$ to form*

the lattice $\Lambda + \mathbf{t}$. Define the subset of all lattice points that cover $B(\mathbf{0}, \kappa_1)$ as $\mathcal{C}(\kappa_1, \kappa_2, \mathbf{t}) \subseteq \Lambda + \mathbf{t}$, i.e.,

$$\mathcal{C}(\kappa_1, \kappa_2, \mathbf{t}) = \{\mathbf{c}_i \in \Lambda + \mathbf{t} : B(\mathbf{c}_i, \kappa_2) \cap B(\mathbf{0}, \kappa_1) \neq \emptyset \\ \text{and } B(\mathbf{0}, \kappa_1) \subseteq \bigcup_i B(\mathbf{c}_i, \kappa_2)\}.$$

Then, the cardinality of $\mathcal{C}$ satisfies:

$$\left(\frac{\kappa_1}{\kappa_2}\right)^m \leq |\mathcal{C}(\kappa_1, \kappa_2, \mathbf{t})| \leq \left(\frac{\kappa_1}{\kappa_2} + 2\right)^m$$

Lemma 19 is a key result which is used to relate the near-optimality dimension of $\mathcal{P}$ and $\mathcal{A}$.

The following lemma provides an upper bound on the number of near-optimal cells in scheme $\mathcal{A}$ relative to scheme $\mathcal{P}$. This relationship is fundamental for comparing the near-optimality dimensions of the two schemes, which will be addressed in a subsequent lemma (Lemma 21).

Lemma 20 *Recollect the definition of the default partitiong scheme $\mathcal{P}$ and $\mathcal{A}$ partitioning scheme from Definition 13. And let $\mathcal{N}_{\mathcal{P}}(\epsilon)$ is the number of cells $\mathcal{P}_{h,i}$ at depth h for which $\sup_{\mathbf{x} \in \mathcal{P}_{h,i}} f(\mathbf{x}) \geq f(\mathbf{x}^*) - \epsilon$. $\mathcal{N}_{\mathcal{A}}(\epsilon), \mathcal{N}_{\mathcal{P}}(\epsilon)$ denote the number of near-optimal cells for $\mathcal{A}, \mathcal{P}$ respectively. Then*

$$\mathcal{N}_{\mathcal{A}}(\epsilon) \leq C\mathcal{N}_{\mathcal{P}}(\epsilon) \quad \textit{where} \quad C = 3^d d^{d-m} (12\sqrt{m})^m.$$

Lemma 21 *For a function in the multi-index class (2) with known $\mathbf{A} \in \mathbb{R}^{m \times d}, m < d$ and let $(\nu, \rho, \eta_{\mathcal{P}}, C)$, $(\nu_{\mathcal{A}}, \rho_{\mathcal{A}}, \eta_{\mathcal{A}}, C_{\mathcal{A}})$ be parameters of $\mathcal{P}, \mathcal{A}$. Let $l_f = f^* - \inf_{\mathbf{x} \in \kappa_1 \mathbb{H}_1^d} f(\mathbf{x})$. Then*

$$\rho_{\mathcal{A}} = \rho^{\beta}, \eta_{\mathcal{A}}(\nu_{\mathcal{A}}, \rho_{\mathcal{A}}, C_{\mathcal{A}}) \leq \eta_{\mathcal{P}}(\nu, \rho, C)/\beta$$

where $\beta = 1 + \frac{d-m}{2m-1}, \tilde{h}_1 = d\lceil \log_3 \kappa_1 \rceil$ and $\kappa_1 = \sqrt{m}\alpha$.

$$C_{\mathcal{A}} = 3^d d^{d-m} (12\sqrt{m})^m C \rho^{-\eta_{\mathcal{P}} \tilde{h}_3}, \nu_{\mathcal{A}} = \max\{\nu, l_f\} \rho^{(1-\beta)(m-1) - \tilde{h}_1}$$

$$\tilde{h}_3 = -\left\lfloor \log_\rho(\max\{\nu, l_f\} \rho^{(1-\beta)(m-1) - \tilde{h}_1}) - \log_\rho(\nu) \right\rfloor \tag{6}$$

When the low rank matrix $\mathbf{A}$ is unknown, we use the estimation guarantees for the learning algorithms of Fornasier et al. [5] and Mousavi-Hosseini et al. [16] that bound the subspace distance $\text{dist}(\mathbf{A}, \hat{\mathbf{A}})$. We can then relate the parameters of the partitioning schemes $\mathcal{A}, \hat{\mathcal{A}}$ obtained using $\mathbf{A}, \hat{\mathbf{A}}$ respectively.

D.1. Using $\hat{\mathcal{A}}$ defined over the estimated $\hat{A}$

We obtain $\hat{\mathbf{A}}$ using a subspace learning algorithm with evaluations of f at selected points in $\mathcal{X}$. We then use $\hat{\mathbf{A}}$ to define a partitioning scheme $\hat{\mathcal{A}}$ (as per Definition 3) and apply `SequOOL` to it. The impact of using an estimated matrix $\hat{\mathbf{A}}$ instead of the true matrix $\mathbf{A}$ in our optimization problem can be quantified using subspace distance. The following is the definition of the subspace distance.

Definition 22 *([3, Lemma 2.5]) Let $\mathbf{A}, \hat{\mathbf{A}} \in \mathbb{R}^{m\times d}$ consists of orthonormal rows such that $\mathbf{A}\mathbf{A}^\top = \mathbf{I}_m$ and $\hat{\mathbf{A}}\hat{\mathbf{A}}^\top = \mathbf{I}_m$. Define two more matrices $\mathbf{A}_\perp, \hat{\mathbf{A}}_\perp \in \mathbb{R}^{(d-m)\times d}$ such that $\begin{bmatrix}\mathbf{A}^\top & \mathbf{A}_\perp^\top\end{bmatrix}$ and $\begin{bmatrix}\hat{\mathbf{A}}^\top & \hat{\mathbf{A}}_\perp^\top\end{bmatrix}$ form two orthonormal bases for $\mathbb{R}^d$. Then the subspace distance between the two row subspaces $(\mathbf{A}, \hat{\mathbf{A}})$ is given by*

$$\operatorname{dist}(\mathbf{A}, \hat{\mathbf{A}}) = \left\|\mathbf{A}^\top\mathbf{A} - \hat{\mathbf{A}}^\top\hat{\mathbf{A}}\right\|_2 = \left\|\hat{\mathbf{A}}_\perp\mathbf{A}^\top\right\|_2 = \left\|\mathbf{A}_\perp\hat{\mathbf{A}}^\top\right\|_2 = \left\|\sin\Theta(\mathbf{A}, \hat{\mathbf{A}})\right\|_2, \tag{7}$$

where $\|\cdot\|_2$ denotes the spectral norm, and $\sin\Theta$ is a diagonal matrix of $\{\sin(\arccos(\sigma_i)) : i = 1, 2, \ldots, m\}$ where σ_i are the singular values of $\hat{\mathbf{A}}\mathbf{A}^\top$ in decreasing order.

Moreover, we have the following equality:

$$\sigma_{min}(\hat{\mathbf{A}}\mathbf{A}^\top) = \cos\theta_m = \sqrt{1-\sin^2\theta_m} = \sqrt{1 - \left\|\sin\Theta(\mathbf{A}, \hat{\mathbf{A}})\right\|_2^2} = \sqrt{1 - \operatorname{dist}^2(\mathbf{A}, \hat{\mathbf{A}})} \tag{8}$$

We observe that assuming $\operatorname{dist}(\mathbf{A}, \hat{\mathbf{A}}) < 1$ implies that $\sigma_{min}(\hat{\mathbf{A}}\mathbf{A}^\top) \neq 0$ and $\operatorname{rank}(\hat{\mathbf{A}}\mathbf{A}^\top) = m$.

Lemma 23 *Consider optimizing a multi-index function $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$ over $\mathcal{X} = [-1, 1]^d$. Let $\mathbf{x}^*$ be an optimizer of f within $\mathcal{X}$, α be as in definition 8 and $\hat{\mathbf{A}}$ be an estimate of $\mathbf{A}$ satisfying $\operatorname{dist}(\mathbf{A}, \hat{\mathbf{A}}) < 1$. Then there exists an $\mathbf{z}^* \in \mathbb{R}^m$ such that $f(\hat{\mathbf{A}}^\top\mathbf{z}^*) = f(\mathbf{x}^*)$ and $\|\mathbf{z}^*\|_\infty \leq \frac{\sqrt{m}}{\sqrt{1-\operatorname{dist}^2(\mathbf{A},\hat{\mathbf{A}})}}\alpha$.*

Given that we only have an estimate $\hat{\mathbf{A}}$ of the true matrix $\mathbf{A}$, and we perform optimization on the subspace spanned by $\hat{\mathbf{A}}$, Lemma 23 guarantees that we can use $\hat{\mathbf{A}}$ and recover f^*. Now, we relate the SequOOL parameters between the partitioning schemes $\mathcal{A}$ and $\hat{\mathcal{A}}$ partitioning schemes.

Lemma 24 *Let $\mathcal{A}$ and $\hat{\mathcal{A}}$ be the partitioning schemes defined in Definition 3. Suppose $\mathcal{A}$ satisfies Assumption 12 with parameters $\nu_{\mathcal{A}}, \rho_{\mathcal{A}}$. Let $l_g = g^* - \inf_{\mathbf{z}\in\kappa_2\alpha\mathbb{H}_1^m} g(\mathbf{A}\hat{\mathbf{A}}^\top\mathbf{z})$. Then $\hat{\mathcal{A}}$ satisfies Assumption 12 with parameters*

$$\nu_{\hat{\mathcal{A}}} = \max\{\nu_{\mathcal{A}}, l_g\}\rho_{\mathcal{A}}^{-\tilde{h}_2}, \rho_{\hat{\mathcal{A}}} = \rho_{\mathcal{A}}$$

where $\tilde{h}_2 = m + m\left\lceil\log_3\frac{2\sqrt{m}\kappa_2}{\kappa_2-1}\right\rceil$ and $\kappa_2 = \frac{\sqrt{m}}{\sqrt{1-\operatorname{dist}^2(\mathbf{A},\hat{\mathbf{A}})}}$.

The above lemma demonstrates that the partitioning scheme $\hat{\mathcal{A}}$ is valid and satisfies Assumption 12 and it establishes the relationship between the SequOOL parameters of the $\hat{\mathcal{A}}$ and $\mathcal{A}$ partitioning schemes.

Algorithm 1 returns the value of $\hat{\alpha}$ based on the Lemma 23. The next lemma relates the near-optimality dimension of $\mathcal{A}$ and $\hat{\mathcal{A}}$.

Lemma 25 *Consider the partitioning scheme $\hat{\mathcal{A}}$ obtained using the estimated $\hat{\mathbf{A}}$. Let $l_g = g^* - \inf_{\mathbf{z}\in\kappa_2\alpha\mathbb{H}_1^m} g(\mathbf{A}\hat{\mathbf{A}}^\top\mathbf{z})$.*

$$\nu_{\hat{\mathcal{A}}} = \max\{\nu_{\mathcal{A}}, l_g\}\rho_{\mathcal{A}}^{-\tilde{h}_2}, \quad \rho_{\hat{\mathcal{A}}} = \rho_{\mathcal{A}}, \quad \eta_{\hat{\mathcal{A}}} \leq \eta_{\mathcal{A}}$$

with

$$C_{\hat{\mathcal{A}}} = C_{\mathcal{A}} 4^m \rho_{\mathcal{A}}^{\eta_{\mathcal{A}} \tilde{h}_4}, \tilde{h}_2 = m + m \left\lceil \log_3 \frac{2\sqrt{m}\kappa_2}{\kappa_2 - 1} \right\rceil, \kappa_2 = \frac{\sqrt{m}}{\sqrt{1 - \mathrm{dist}^2(\mathbf{A}, \hat{\mathbf{A}})}}$$

$$\tilde{h}_4 = -\left\lfloor \log_{\rho_{\mathcal{A}}}(\max\{\nu_{\mathcal{A}}, l_g\}\rho_{\mathcal{A}}^{-\tilde{h}_2}) - \log_{\rho_{\mathcal{A}}}(\nu_{\mathcal{A}}) \right\rfloor \tag{9}$$

This corollary synthesizes the relationships established in the preceding lemmas, providing a direct comparison between the estimated scheme $\hat{\mathcal{A}}$ and the default partitioning scheme $\mathcal{P}$. It combines the two-step process of relating $\mathcal{A}$ to $\mathcal{P}$ and then $\hat{\mathcal{A}}$ to $\mathcal{A}$, yielding a comprehensive set of relationships for the SequOOL parameters, near-optimality dimensions, and associated constants.

Corollary 26 *Referring to Lemma 21, for the partitioning scheme $\mathcal{P}$ with $\kappa = 1$, we have*

$$\rho_{\mathcal{A}} = \rho^{\beta}, \nu_{\mathcal{A}} = \max\{\nu, l_f\}\rho^{(1-\beta)(m-1)-\tilde{h}_1}, \eta_{\mathcal{A}}(\nu_{\mathcal{A}}, \rho_{\mathcal{A}}, C_{\mathcal{A}}) \leq \frac{\eta_{\mathcal{P}}(\nu, \rho, C)}{\beta}$$

and $C_{\mathcal{A}} = 3^d d^{d-m}(12\sqrt{m})^m C\rho^{-\eta_{\mathcal{P}}\tilde{h}_3}$. By utilizing Lemma 24 and Lemma 25, we establish the following relationships among the parameters associated with $\hat{\mathcal{A}}$ and $\mathcal{P}$.

$$\rho_{\hat{\mathcal{A}}} = \rho^{\beta}, \nu_{\hat{\mathcal{A}}} = \max\{\max\{\nu, l_f\}\rho^{(1-\beta)(m-1)-\tilde{h}_1}, l_g\}\rho_{\mathcal{A}}^{-\tilde{h}_2},$$
$$\eta_{\hat{\mathcal{A}}}(\nu_{\hat{\mathcal{A}}}, \rho_{\hat{\mathcal{A}}}, C_{\hat{\mathcal{A}}}) \leq \frac{\eta_{\mathcal{P}}(\nu, \rho, C)}{\beta}, C_{\hat{\mathcal{A}}} = 3^d d^{d-m}(12\sqrt{m})^m C\rho^{-\eta_{\mathcal{P}}\tilde{h}_3} 4^m \rho^{\eta_{\mathcal{P}}\tilde{h}_4}.$$

with $\tilde{h}_2 = m + m\left\lceil \log_3 \frac{2\sqrt{m}\kappa_2}{\kappa_2 - 1} \right\rceil$ and $\tilde{h}_1 = d\lceil \log_3 \sqrt{m}\alpha \rceil$ with $\kappa_2 = \frac{\sqrt{m}}{\sqrt{1-\mathrm{dist}^2(\mathbf{A}, \hat{\mathbf{A}})}}$

$$\tilde{h}_3 = -\left\lfloor \log_{\rho}(\max\{\nu, l_f\}\rho^{(1-\beta)(m-1)-\tilde{h}_1}) - \log_{\rho}(\nu) \right\rfloor$$
$$\tilde{h}_4 = -\left\lfloor \log_{\rho_{\mathcal{A}}}(\max\{\nu_{\mathcal{A}}, l_g\}\rho_{\mathcal{A}}^{-\tilde{h}_2}) - \log_{\rho_{\mathcal{A}}}(\nu_{\mathcal{A}}) \right\rfloor$$

Theorem 27 *For a function in the multi-index class (2), the regret of SequOOL applied on the partitioning scheme using $\hat{\mathbf{A}}$ returned by Algorithm 1 and $\hat{\alpha} = \sqrt{dm}/\sqrt{1 - \mathrm{dist}^2(\mathbf{A}, \hat{\mathbf{A}})}$ satisfies*

- *If $\eta_{\mathcal{P}} = 0, r_n \leq \gamma(\nu, \rho)\rho^{-\beta\tilde{h}_2}\rho^{\frac{\beta}{C_1}\lfloor \frac{n}{\log n} \rfloor}$*
- *If $\eta_{\mathcal{P}} > 0, r_n \leq \gamma(\nu, \rho)\rho^{-\beta\tilde{h}_2}\left(\frac{\tilde{n}}{\log \tilde{n}}\right)^{-\frac{\beta}{\eta_{\mathcal{P}}}}$*

where, $\gamma(\nu, \rho) = \max\{\max\{\nu, l_f\}\rho^{(1-\beta)(m-1)-\tilde{h}_1}, l_g\}, C_1 = 3^d d^{d-m}(12\sqrt{m})^m C\rho^{-\eta_{\mathcal{P}}\tilde{h}_3} 4^m \rho^{\eta_{\mathcal{P}}\tilde{h}_4}$

$$\tilde{n} = \lfloor n/\overline{\log} n \rfloor \eta_{\mathcal{P}} \log(1/\rho)/C_1$$

Where $\tilde{h}_1, \tilde{h}_2$ are defined in Lemma 21 and Lemma 25. $\tilde{h}_3, \tilde{h}_4$ are from equations 6 and 9 respectively.

When $\eta_{\mathcal{P}} > 0$, Algorithm 1 has $r_n = \tilde{O}(n^{-\beta/\eta_{\mathcal{P}}}) = \tilde{O}(n^{-(1+\frac{d-m}{2m-1})/\eta_{\mathcal{P}}})$ while default SequOOL would give $\tilde{O}(n^{-1/\eta_{\mathcal{P}}})$, showing our approach reduces the regret at a faster rate. The proof of this theorem follows from applying the SequOOL parameters derived in Corollary 26 to Theorem 5 of Bartlett et al. [1]. Detailed proof is in Appendix F.19.

[16], controls the closeness between true and estimated subspace expressed in terms of $\|\mathbf{W}_\perp\|_F$. However, for out Algorithm 2, we require an upper bound on the distance between the true subspace $\mathbf{A}$ and its estimate $\hat{\mathbf{A}}$. This lemma bridges this gap by providing an upper bound on $\operatorname{dist}(\mathbf{A}, \hat{\mathbf{A}})$ in terms of $\|\mathbf{W}_\perp\|_F$ and the singular values of $\mathbf{W}$.

Lemma 28 *Given $\mathbf{A} \in \mathbb{R}^{m\times d}$ satisfying $\mathbf{A}\mathbf{A}^\top = \mathbf{I}_m$, let $\mathbf{W} \in \mathbb{R}^{p\times d}$ be any matrix such that $\operatorname{rank}(\mathbf{W}) \geq m$ and $d \geq m$. Consider the singular value decomposition of $\mathbf{W} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$ and collect the top m right singular vectors in the matrix $\hat{\mathbf{A}} = \begin{bmatrix}\mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_m\end{bmatrix}^\top$, where $\mathbf{v}_i$ is the ith column of $\mathbf{V}$. Recollect the definition of $\sin\Theta(\mathbf{A}, \hat{\mathbf{A}})$ from the Definition 22. Using $\|\cdot\|_F$ to denote Frobenius norm and σ_m to denote the mth singular value of $\mathbf{W}$, we have that*

$$\operatorname{dist}(\mathbf{A}, \hat{\mathbf{A}}) = \left\|\sin\Theta(\mathbf{A}, \hat{\mathbf{A}})\right\|_2 \leq \frac{\|\mathbf{W}_\perp\|_F}{\sigma_m}. \tag{10}$$

D.2. Supporting lemmas

The following lemmas and assumptions are needed to obtain guarantees on learning a good estimate $\hat{\mathbf{A}}$.

Assumption 29 *[16] The student model is a two-layer neural network eq. (3) trained over the data set $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i\geq 1}$, where the target values $y^{(i)}$ are generated according to the teacher model eq. (2) and the inputs satisfy $x^{(i)} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. The link function $g(\cdot)$ is weakly differentiable.*

Assumption 30 *[16] For all $1 \leq i \leq m, 1 \leq j \leq d$, we initialize the NN weights and biases with $\sqrt{d}\mathbf{W}_{ij}^0 \overset{iid}{\sim} \mathcal{N}(0,1), ma_i^0 \overset{iid}{\sim} \operatorname{Unif}([-1,1])$, and $b_i^0 \overset{iid}{\sim} \operatorname{Unif}(\{-1,1\})$*

Lemma 31 ([16], Theorem 3) *Consider running T SGD iterations over samples satisfying theorem 29, with an initialization satisfying theorem 30, and using the following decaying step size schedule. Assuming ReLU non-linearity, let $\zeta := 2\sqrt{2/e\pi}$. Choose the decreasing step size $\eta_t = m\frac{2(t+t^*)+1}{\gamma(t+t^*+1)^2}, \tilde{\lambda} \geq \gamma + \zeta$ and $t^* \asymp \frac{\tilde{\lambda}}{\gamma}$ for any $\gamma > 0$. Then, for $\lambda = \frac{\tilde{\lambda}}{m}$, with probability at least $1-\delta$,*

$$\frac{\|\mathbf{W}_\perp^\top\|_F}{\sqrt{m}} \lesssim \sqrt{\frac{(d+\log(1/\delta))}{\gamma^2 T}}$$

whenever $m \gtrsim \log(1/\delta)$ and $T \gtrsim \frac{\tilde{\lambda}^2}{d+\log(1/\delta)}$.

The following Lemma is to control the subspace distance using compressed sensing algorithm.

Theorem 32 ([5], Theorem 4.1)

Let $f(\mathbf{x}) = g(\mathbf{A}\mathbf{x})$ be a function where $\mathbf{A}$ is a $k \times d$ matrix with orthonormal rows, and g is a twice continuously differentiable function. Assume that $\mathbf{H}^f = \mathbf{A}^\top\mathbf{H}_g\mathbf{A}$ is well-conditioned with

$\sigma_k(\mathbf{H}^f) \geq \alpha > 0$. Let $\hat{\mathbf{A}}$ be the matrix obtained from the Dantzig Selector approximation $\hat{\mathbf{X}}$ of the matrix $\mathbf{X}$ of gradients of f at m_X random points. Then, with high probability, the distance between the subspaces spanned by the rows of $\mathbf{A}$ and $\hat{\mathbf{A}}$ is bounded by:

$$\left\|\mathbf{A}^\top\mathbf{A} - \hat{\mathbf{A}}^\top\hat{\mathbf{A}}\right\|_F \leq \frac{2\nu_2}{\sqrt{\alpha(1-s)} - \nu_2}$$

where

$$\nu_2 = Ck^{1/q}\left(\frac{m_\Phi}{\log(d/m_\Phi)}\right)^{1/2-1/q} + \frac{\epsilon k^2}{\sqrt{m_\Phi}}$$

and C is a constant depending on the parameters C_1 and C_2 from the conditions on $\mathbf{A}$ and g, m_Φ is the number of derivative directions, ϵ is the step size used in the finite difference approximation, d is the ambient dimension, and $s \in (0,1)$ is a parameter.

Appendix E. Illustrative experiments to motivate lookahead direction selection

Using a partitioning scheme with a lower near-optimality dimension can lead to a faster decrease in regret. Empirically, we observe that the regret for SequOOL applied for a budget of 200 evaluations of the function in the Example 6 was 5×10^{-10} for the default partitioning scheme and 5.8×10^{-12} for the direction selection strategy in Example (6). This example indicates that it can be beneficial to use a direction splitting strategy that adapts to the function being optimized.

Additionally, we evaluate the regret for different choices of $\mathbf{A}$, by parameterizing $\mathbf{A} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$ and changing the rotation angle from 0 to $\pi/8$ while keeping the direction selection strategy the same as in Example (6). Figure 3 shows that the regret varies significantly over the range of angles. This shows that minimizing the angle of discrepancy between $\mathbf{A}$ and the true directions of variation (which are the standard x_1 and x_2 axes in this example) is beneficial in reducing the regret.

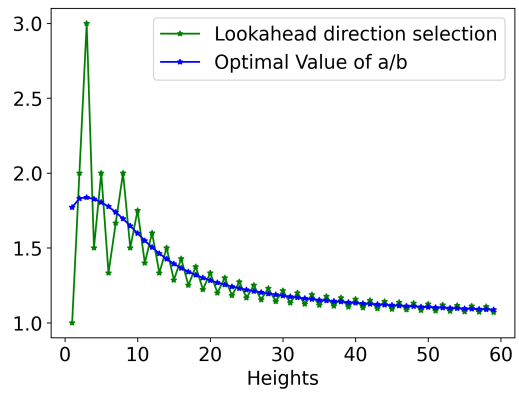

Figure 2: Optimal splitting ratio between the first and second directions. Green curve is obtained using the strategy in Definition 33

Given a particular $\mathbf{A}$ and function f, we consider the question of identifying an appropriate direction selection strategy that minimizes $\mathcal{N}_\mathcal{A}(h)$ at all heights. The two example strategies for $f(x_1, x_2) = 1 - |x_1| - x_2^2$ we have seen so far are the default round-robin (equivalently 1:1) splitting and the 2:1 splitting in Example(6). For an $\mathbf{A}$ with $\theta = \pi/48$, we minimize the number of near-optimal cells $\mathcal{N}_\mathcal{A}$ at different heights by choosing the best splitting ratio at each height and plot the ratios as the blue line in Figure 2. We see that as the height increases to infinity the optimal split ratio converges to 1. However, at lower heights, the optimal split ratio is greater than 1 and takes its maximum value 1.83 at $h = 3$.

Definition 33 *Lookahead strategy for direction selection. Given an estimated $\hat{f}$ and the current tree of partitions till depth h, the lookahead strategy first evaluates the different number of near-optimal*

cells for $\hat{f}$ at depth $h+1$ by splitting along each of the m directions. It then greedily selects the direction to be split at $h+1$ as the direction that results in the lowest number of near-optimal cells.

In a numerical experiment, we see that the lookahead strategy closely matches the optimal split ratio (shown as green points in Fig 2) over all heights. This also results in it having a low regret than the default partitioning scheme. Empirically, we observe that the regret for SequOOL applied for a budget of 500 evaluations of the function in the previous lemma was 5.68×10^{-10} for the partitioning scheme $\mathcal{A}$ using the lookahead strategy for direction selection, 1.98×10^{-5} for $\mathcal{A}$ with 1:1 splitting, and 1.8×10^{-4} for $\mathcal{A}$ with the 2:1 splitting strategy from Example (6).

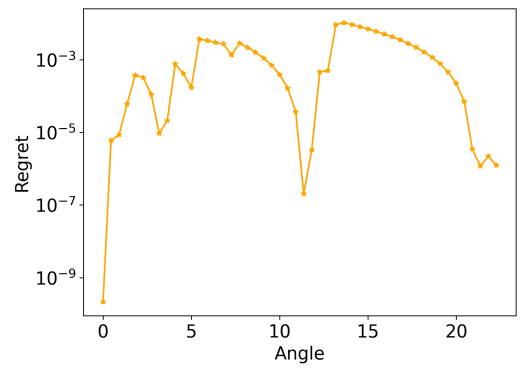

Figure 3: Regret at $n = 300$ for SequOOL on $\mathbf{A}$ with varying θ.

Appendix F. Proofs

Some of the facts which we use in our proofs.

Key inequalities for the matrix norms include: $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1$. Holder's inequality, applicable for $p, q \geq 1$ where $\frac{1}{p} + \frac{1}{q} = 1$, states that $|\mathbf{x}^\top \mathbf{y}| \leq \|\mathbf{x}\|_p \|\mathbf{y}\|_q$. The triangle inequality, valid for any $p \geq 1$, asserts that $\|\mathbf{x} + \mathbf{y}\|_p \leq \|\mathbf{x}\|_p + \|\mathbf{y}\|_p$.

Matrix Norms: For a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the operator norm is:

$$\|\mathbf{A}\|_p = \sup_{\mathbf{x} \neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$$

When $p = \infty$, $\|\mathbf{A}\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^n |a_{ij}|$ and when $p = 2$, $\|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A})$, Where $\sigma_{\max}(\mathbf{A})$ represents the largest singular value of matrix $\mathbf{A}$. Additionally, the Frobenius norm is given by $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2}$.

For a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ of rank r, the following inequalities hold:

$$\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \sqrt{r}\, \|\mathbf{A}\|_2 \,, \quad \frac{1}{\sqrt{n}} \|\mathbf{A}\|_\infty \leq \|\mathbf{A}\|_2 \leq \sqrt{m}\, \|\mathbf{A}\|_\infty \tag{11}$$

For the operator norm $\|\cdot\|_2$, one has

$$\|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_2 \,, \quad \|\mathbf{AB}\|_2 \geq \|\mathbf{A}\|_2 \sigma_{\min}(\mathbf{B}), \quad \|\mathbf{AB}\|_2 \geq \|\mathbf{B}\|_2 \sigma_{\min}(\mathbf{A}) \tag{12}$$

Suppose a matrix $\mathbf{A}$ consists of orthonormal rows or columns, then $\|\mathbf{A}\|_2 = \left\|\mathbf{A}^\top\right\|_2 = 1$

F.1. Proof of Example 4

Proof Consider the function $f(x_1, x_2) = g(\mathbf{A}\mathbf{x}) = 1 - |x_1|$ with $\mathbf{A} = [1, 0]$ and $g(z) = 1 - |z|$. Let $\mathcal{P}, \mathcal{A}$ be the partitioning schemes defined in theorem 2 with $\kappa = 1$ and parameters $(\nu, \rho), (\nu_\mathcal{A}, \rho_\mathcal{A})$ respectively. For the $\mathcal{P}$ partitioning scheme, along the X axis, the side lengths of the children at depth

h are given by $3^{-\lceil h/2 \rceil}$ and along the Y axis, it is $3^{-\lfloor h/2 \rfloor}$. Consider $f(x_1^*, x_2^*) - f(x_1, x_2) = |x_1|$, and the cell with the representative origin is the $\mathcal{P}_h^*$ cell, therefore

$$|x_1| = 3^{-\lceil h/2 \rceil} \leq 3^{-h/2} \leq (1/\sqrt{3})^h.$$

According to Assumption 12, the appropriate values are $\nu = 1$, $\rho = 1/\sqrt{3}$. Now consider a rectangle region R with corners $\{(-3\nu\rho^h, -1), (3\nu\rho^h, -1), (3\nu\rho^h, 1), (-3\nu\rho^h, 1)\}$. Since $f^* - f(x_1, x_2) = |x_1|, \forall (x_1, x_2) \in R, f^* - f(x_1, x_2) \leq 3\nu\rho^h$. Thus any cell in $\mathcal{P}$ that has a non-empty intersection with R is a near-optimal cell. Each cell at depth h has an area of $3^{-\lceil h/2 \rceil} 3^{-\lfloor h/2 \rfloor}$, therefore

$$\mathcal{N}_{\mathcal{P}}(h) \geq \frac{\text{Area}(R)}{\text{Area}(\mathcal{P}(h,i))} = \frac{4(3(1/\sqrt{3})^h)}{3^{-\lceil h/2 \rceil} 3^{-\lfloor h/2 \rfloor}},$$

yielding that since $\mathcal{N}_{\mathcal{P}}(h) = \Omega(\rho^{-h})$, from the Definition 13 we get $\eta_{\mathcal{P}} = 1$.

For the $\mathcal{A}$ partitioning scheme, we first note that $\alpha = 1$. Along X axis, the side length of the children at depth h is 3^{-h}, therefore

$$f(x_1^*, x_2^*) - f(x_1, x_2) = |x_1| = 3^{-h} \leq (1/3)^h,$$

yielding the values are $\nu_{\mathcal{A}} = 1$, $\rho_{\mathcal{A}} = 1/3$. Now consider a line segment L with endpoints $\{(-3\nu_{\mathcal{A}}\rho_{\mathcal{A}}^h, 0), (3\nu_{\mathcal{A}}\rho_{\mathcal{A}}^h, 0)\}$. Since $f^* - f(x_1, x_2) = |x_1|, \forall (x_1, 0) \in L, f^* - f(x_1, x_2) \leq 3\nu_{\mathcal{A}}\rho_{\mathcal{A}}^h$. Thus any cell in $\mathcal{A}$ that has an intersection with the L is a near-optimal cell. Every $\mathcal{A}(h,i)$ cell at depth h has a length of 3^{-h}, therefore,

$$\mathcal{N}_{\mathcal{A}}(3\nu_{\mathcal{A}}\rho_{\mathcal{A}}^h) \leq 2 + \frac{\text{len}(L)}{\text{len}(\mathcal{A}(h,i))} = 2 + \frac{2(3(1/3)^h)}{3^{-h}} = 8,$$

where the additional term 2 accounts for cells in $\mathcal{A}$ that partially intersect L at its endpoints. Hence $\mathcal{N}_{\mathcal{A}}(3\nu_{\mathcal{A}}\rho_{\mathcal{A}}^h)$ is a constant and $\eta_{\mathcal{A}} = 0$. ■

F.2. Proof of Example 10

Proof Consider the function $g(k) = 1 - |k|$ with $\mathbf{A} = [1, 1]$. Then, $f(x_1, x_2) = g(\mathbf{Ax}) = 1 - |x_1 + x_2|$. For the $\mathcal{P}$ partitioning scheme with $\kappa = 1$, along the X axis, the side lengths of the children at depth h along the X, Y axes are $3^{-\lceil h/2 \rceil}, 3^{-\lfloor h/2 \rfloor}$ respectively. Consider $f(x_1^*, x_2^*) - f(x_1, x_2) = |x_1 + x_2|$, and the cell with the representative origin is the $\mathcal{P}_h^*$ cell, therefore

$$|x_1 + x_2| = 3^{-\lceil h/2 \rceil} + 3^{-\lfloor h/2 \rfloor} \leq 2 \cdot 3^{-(h-1)/2} \leq 2\sqrt{3}(1/\sqrt{3})^h.$$

According to Assumption 12, $\nu = 2\sqrt{3}$, $\rho = 1/\sqrt{3}$. Consider a rhombus region T formed by coordinates $\{(-3\nu\rho^h, 0), (0, -3\nu\rho^h), (3\nu\rho^h, 0), (0, 3\nu\rho^h)\}$. Since $f^* - f(x_1, x_2) = |x_1 + x_2|$, $\forall (x_1, x_2) \in T, f^* - f(x_1, x_2) \leq 3\nu\rho^h$. This says, any cell in $\mathcal{P}$ that has an intersection with the T is near-optimal cell. Each cell at depth h has an area of $3^{-\lceil h/2 \rceil} 3^{-\lfloor h/2 \rfloor}$, therefore,

$$\mathcal{N}_{\mathcal{P}}(3\nu\rho^h) \leq \left(1 + \frac{3\nu\rho^h}{3^{-\lceil h/2 \rceil}}\right)\left(1 + \frac{3\nu\rho^h}{3^{-\lfloor h/2 \rfloor}}\right) \leq O(1)$$

Hence $\mathcal{N}_{\mathcal{P}}(3\nu\rho^h)$ is independent of h and $\eta_{\mathcal{P}} = 0$.

For the $\mathcal{A}$ partitioning scheme, along the X axis and Y axis, the side lengths of the children at depth h is given by 3^{-h} and 3^{-h}, giving that

$$f(x_1^*, x_2^*) - f(x_1, x_2) = |x_1 + x_2| = 2 \cdot 3^{-h} \leq 2(1/3)^h.$$

According to Assumption 12, the values are $\nu_{\mathcal{A}} = 2$, $\rho_{\mathcal{A}} = 1/3$.

Further, we see from Definition 8 that $\alpha = \sqrt{2}$ in this example. ■

F.3. Proof of Example 6

Proof

For the $\mathcal{P}$ partitioning scheme, the side lengths of the children at depth h are as follows: along the X axis, $3^{-\lceil h/2 \rceil}$. Along the Y axis, it is $3^{-\lfloor h/2 \rfloor}$. Consider $f(x_1^*, x_2^*) - f(x_1, x_2) = |x_1| + x_2^2$, and the cell with the representative origin is the $\mathcal{P}_h^*$ cell, therefore

$$\begin{aligned} |x_1| + x_2^2 &= 3^{-\lceil h/2 \rceil} + 3^{-2\lfloor h/2 \rfloor} \leq 3^{-\lfloor h/2 \rfloor} + 3^{-2\lfloor h/2 \rfloor} \\ &\leq 2 \cdot 3^{-\lfloor h/2 \rfloor} \leq 2 \cdot 3^{-(h-1)/2} \leq 2\sqrt{3}(1/\sqrt{3})^h \end{aligned}$$

According to Assumption 12, the appropriate values are $\nu = 2\sqrt{3}$, $\rho = 1/\sqrt{3}$ and $\nu\rho^h = 2\sqrt{3}(1/\sqrt{3})^h$.

Consider a region R which is given by $|x_1| + x_2^2 \leq 3\nu\rho^h$. Since $f^* - f(x_1, x_2) = |x_1| + x_2^2$, $\forall (x_1, x_2) \in R, f^* - f(x_1, x_2) \leq 3\nu\rho^h$. Thus any cell in $\mathcal{P}$ that has a non-empty intersection with R is a near-optimal cell. Each cell at depth h has an area of $3^{-\lceil h/2 \rceil}3^{-\lfloor h/2 \rfloor}$, therefore

$$\mathcal{N}_{\mathcal{P}}(h) \geq \frac{\text{Area}(R)}{\text{Area}(\mathcal{P}(h,i))}$$

Now, we compute the area of the region formed by the curve $|x_1| + x_2^2 = 3\nu\rho^h$ which is given by

$$4\int_0^{\sqrt{3\nu\rho^h}} \int_0^{3\nu\rho^h - x_2^2} dx_1 dx_2 = {}^8/_3(3\nu\rho^h)^{3/2}$$

Thus,

$$\mathcal{N}_{\mathcal{P}}(h) \geq \frac{\text{Area}(R)}{\text{Area}(\mathcal{P}(h,i))} = \frac{8/3(6\sqrt{3}(\frac{1}{\sqrt{3}})^h)^{3/2}}{3^{-\lceil h/2 \rceil}3^{-\lfloor h/2 \rfloor}} = \Omega\left((\frac{1}{\sqrt{3}})^{-h/2}\right)$$

Hence, $\eta_{\mathcal{P}} \geq 0.5$

The $\mathcal{P}_h^*$ cell contains the point $(0,0)$. Since $\mathcal{P}$ is an axis-aligned partitioning scheme, we can bound the number of cells directly above, i.e., having the same x coordinate of their representative as that of, $\mathcal{P}_h^*$ by the value $\sqrt{3\nu\rho^h}/3^{-\lfloor h/2 \rfloor}$. In a similar manner, we can bound the maximum number of cells having their representative's y coordinate to be the same as that of $\mathcal{P}_h^*$ by $3\nu\rho^h/3^{-\lceil h/2 \rceil}$. Since $\mathcal{P}$ is an axis aligned partitioning scheme, the previous two bounds imply that number of near

optimal cells are upper bounded by the product of number of cells along x axis times number of cells along y axis. Thus,

$$\mathcal{N}_{\mathcal{P}}(h) \leq \left(1 + \left\lceil \frac{\sqrt{3 \cdot 2\sqrt{3}(1/\sqrt{3})^h}}{3^{-\lceil h/2 \rceil}} \right\rceil\right) \left(1 + \left\lceil \frac{3 \cdot 2\sqrt{3}(1/\sqrt{3})^h}{3^{-\lfloor h/2 \rfloor}} \right\rceil\right) = O\left((\frac{1}{\sqrt{3}})^{-h/2}\right)$$

Thus, $\eta_{\mathcal{P}} \leq 0.5$, hence $\eta_{\mathcal{P}} = 0.5$

For the $\mathcal{A}$ partitioning scheme, the side lengths of the children at depth h are as follows: along the X axis, $3^{-(h+1)}$ if h is odd; otherwise, 3^{-h}. Along the Y axis, it is $3^{-\lfloor h/2 \rfloor}$. Let h is even. consider $f(x_1^*, x_2^*) - f(x_1, x_2) = |x_1| + x_2^2$ and the cell with the representative origin is the $\mathcal{A}_h^*$ cell, therefore

$$|x_1| + x_2^2 = 3^{-h} + 3^{-2\lfloor h/2 \rfloor} \leq 3^{-h} + 3^{-2(h-1)/2} \leq 4 \cdot 3^{-h}$$

According to Assumption 12, we have $\nu_{\mathcal{A}} = 4$, $\rho_{\mathcal{A}} = 1/3$, and $\nu_{\mathcal{A}}\rho_{\mathcal{A}}^h = 4 \cdot 3^{-h}$.

The $\mathcal{A}_h^*$ cell contains the point $(0, 0)$. Since $\mathcal{A}$ is an axis-aligned partitioning scheme, we can bound the number of cells directly above, i.e., having the same x coordinate of their representative as that of, $\mathcal{A}_h^*$ by the value $\sqrt{3\nu\rho^h}/3^{-h/2}$. In a similar manner, we can bound the maximum number of cells having their representative's y coordinate to be the same as that of $\mathcal{P}_h^*$ by $3\nu\rho^h/3^{-h}$. Since $\mathcal{P}$ is an axis aligned partitioning scheme, the previous two bounds imply that number of near optimal cells are upper bounded by the product of number of cells along x axis times number of cells along y axis. Thus,

$$\mathcal{N}_{\mathcal{A}}(\nu_{\mathcal{A}}\rho_{\mathcal{A}}^h) \leq \left(1 + \left\lceil \frac{\sqrt{3 \cdot 4 \cdot 3^{-h}}}{3^{-h/2}} \right\rceil\right) \left(1 + \left\lceil \frac{3 \cdot 4 \cdot 3^{-h}}{3^{-h}} \right\rceil\right) = 65$$

When h is odd, following the same steps will give $\mathcal{N}_{\mathcal{A}}(\nu_{\mathcal{A}}\rho_{\mathcal{A}}^h) \leq 148$.

Hence, $\mathcal{N}_{\mathcal{A}}(\nu_{\mathcal{A}}\rho_{\mathcal{A}}^h)$ is a constant and $\eta_{\mathcal{A}} = 0$. ■

F.4. Proof of Proposition 9

Proof

We claim that $\mathbf{z}^* = \mathbf{A}\mathbf{x}^*$, where $\mathbf{x}^* \in \mathcal{X}$ is the optimizer of f. This choice of $\mathbf{z}^*$ satisfies $f(\mathbf{A}^\top \mathbf{z}^*) = f(\mathbf{x}^*)$ as required in the Lemma. we will now show that the infinity norm of this $\mathbf{z}^*$ is less than or equal to α.

First, observe that for any $\mathbf{x} \in [-1, 1]^d$, we can express $\mathbf{x}$ as a convex combination of corner points:

$$\mathbf{x} = \sum_{j=1}^{2^d} \mathbf{c}_j \alpha_j, \quad \alpha_j \geq 0, \sum_{j=1}^{2^d} \alpha_j = 1 \tag{13}$$

Now, let's consider the infinity norm of $\mathbf{Ax^*}$:

$$\begin{aligned}
\|\mathbf{Ax^*}\|_\infty &= \left\| \sum_{j=1}^{2^d} \mathbf{Ac}_j \alpha_j \right\|_\infty && \text{(Using Equation 13)} \\
&\leq \sum_{j=1}^{2^d} \|\mathbf{Ac}_j \alpha_j\|_\infty && \text{(Triangle Inequality of Norms)} \\
&= \sum_{j=1}^{2^d} |\alpha_j| \, \|\mathbf{Ac}_j\|_\infty && \text{(Absolute Homogeneity property of Norms)} \\
&\leq \alpha \sum_{j=1}^{2^d} |\alpha_j| \leq \alpha && \text{(From the definition of } \alpha \text{ in the Proposition statement)}
\end{aligned}$$

Therefore, we have shown that $\|\mathbf{z}^*\|_\infty = \|\mathbf{Ax^*}\|_\infty \leq \alpha$, which completes the proof. ■

F.5. Proof of Proposition 14

Proof

Let $\mathbf{x}_h^*$ be the representative point (midpoint) of the $\mathcal{P}_h^*$ cell and $\mathbf{x}$ is within the $\mathcal{P}_h^*$ cell.

Let $\mathbf{x}$ be a point in a hyperrectangle centered at $\mathbf{x}_h^*$ with side lengths $2\kappa\mathbf{c}$, where $\mathbf{c}$ is a vector of side lengths. Then, this point $\mathbf{x}$ satisfies $|x_i - x_{h,i}^*| \leq \kappa c_i$ for all $i \in \{1, \ldots, d\}$. Equivalently, this set of inequalities can be written as $\max_{i\in\{1,\ldots,d\}} \left|\frac{x_i - x_{h,i}^*}{c_i}\right| \leq \kappa$. Using the infinity norm, we can concisely express this conditions as $\left\|\frac{\mathbf{x}-\mathbf{x}_h^*}{\mathbf{c}}\right\|_\infty \leq \kappa$, where the division is performed element-wise.

For the partition scheme, we perform trisection along each axis in a round-robin manner. After h iterations, the side lengths are given by:

$$c_i = \kappa 3^{-\left\lfloor \frac{h+d-i}{d} \right\rfloor}, \quad i \in \{1, \ldots, d\} \tag{14}$$

where $\left\lfloor \frac{h+d-i}{d} \right\rfloor$ represents the number of trisections applied to dimension i. Therefore, the cell $\mathcal{P}_h^*$ can be described as:

$$\mathcal{P}_h^* = \left\{ \mathbf{x} \in \mathbb{R}^d : \left\| \frac{\mathbf{x} - \mathbf{x}_h^*}{\mathbf{c}} \right\|_\infty \leq \kappa \quad \text{with} \quad \mathbf{c} = [3^{-\left\lfloor \frac{h+d-i}{d} \right\rfloor}]_{i=1}^d. \right\} \tag{15}$$

From the theorem 3, we have $\mathcal{A}_h^* \triangleq \{\mathbf{A}^\top \boldsymbol{\alpha} : \boldsymbol{\alpha} \in \mathcal{T}_h^*\}$ and

$$\mathcal{T}_h^* = \left\{ \boldsymbol{\alpha} \in \mathbb{R}^m : \left\| \frac{\boldsymbol{\alpha} - \boldsymbol{\alpha}_h^*}{\mathbf{s}} \right\|_\infty \leq \alpha \quad \text{with} \quad \mathbf{s} = [3^{-\left\lfloor \frac{h+m-i}{m} \right\rfloor}]_{i=1}^d. \right\} \tag{16}$$

Hence,

$$\mathcal{A}_h^* = \left\{ \mathbf{A}^\top \boldsymbol{\alpha} : \boldsymbol{\alpha} \in \mathbb{R}^m, \left\| \frac{\boldsymbol{\alpha} - \boldsymbol{\alpha}_h^*}{\mathbf{s}} \right\|_\infty \leq \alpha \quad \text{with} \quad \mathbf{s} = [3^{-\left\lfloor \frac{h+m-i}{m} \right\rfloor}]_{i=1}^d. \right\} \tag{17}$$

■

F.6. Proof of Lemma 15

Proof Following the Definition 2, consider the partitioning schemes $\mathcal{P}$ and $\mathcal{A}$.

According to Proposition 14, we can express the cell $\mathcal{A}_h^*$ as:

$$\mathcal{A}_h^* = \{\mathbf{A}^\top \boldsymbol{\alpha} : \boldsymbol{\alpha} \in \mathbb{R}^m, \left\| \frac{\boldsymbol{\alpha} - \boldsymbol{\alpha}_h^*}{\mathbf{s}} \right\|_\infty \leq \alpha \quad \text{with} \quad \mathbf{s} = [3^{-\lfloor \frac{h+m-i}{m} \rfloor}]_{i=1}^m\}$$

At depth $h = km + i$, for $i \in [1 : m-1]$ and $k \in \mathbb{N}_0$, the side lengths simplifies to: $\mathbf{s} = [3^{-\mathbb{1}(j \leq i)-k}]_{j=1}^m$ Similarly,

$$\mathcal{P}_h^* = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^d, \left\| \frac{\mathbf{x} - \mathbf{x}_h^*}{\mathbf{c}} \right\|_\infty \leq 1\} \quad \text{with} \quad \mathbf{c} = [3^{-\lfloor \frac{h+d-i}{d} \rfloor}]_{i=1}^d.$$

At depth $h = kd + i$, for $i \in [1 : d-1]$ and $k \in \mathbb{N}_0$, the side lengths for $\mathcal{P}_h^*$ become: $\mathbf{c} = [3^{-\mathbb{1}(j \leq i)-k}]_{j=1}^d$.

And let us consider another partitioning scheme $\mathcal{G} = \kappa\mathcal{P}$. Let us denote $\tilde{\mathbf{x}}_h^*$ to be the representative of the $\mathcal{G}_h^*$ cell. Using Proposition 14, the cell $\mathcal{G}_h^*$ can be written as

$$\mathcal{G}_{kd+i}^* = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^d, \left\| \frac{\mathbf{x} - \tilde{\mathbf{x}}_{kd+i}^*}{\mathbf{c}} \right\|_\infty \leq \kappa\} \quad \text{with} \quad \mathbf{c} = [3^{-\mathbb{1}(j \leq i)-k}]_{j=1}^d$$

Consider an element $\mathbf{x} = \mathbf{A}^\top \boldsymbol{\alpha} \in \mathcal{A}_{km+i}^*$. we now proceed with the following chain of inequalities:

$$\begin{aligned}
\left\|\mathbf{x} - \tilde{\mathbf{x}}_{kd+i}^*\right\|_\infty &\leq \|\mathbf{x} - \mathbf{x}^*\|_\infty + \left\|\mathbf{x}^* - \tilde{\mathbf{x}}_{kd+i}^*\right\|_\infty && \text{(Vector Norm property)} \\
&= \left\|\mathbf{A}^\top \boldsymbol{\alpha} - \mathbf{A}^\top \boldsymbol{\alpha}^*\right\|_\infty + \left\|\mathbf{x}^* - \tilde{\mathbf{x}}_{kd+i}^*\right\|_\infty && (\mathbf{x}^* = \mathbf{A}^\top \boldsymbol{\alpha}^*) \\
&= \left\|\mathbf{A}^\top (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)\right\|_\infty + \left\|\mathbf{x}^* - \tilde{\mathbf{x}}_{kd+i}^*\right\|_\infty \\
&\leq \left\|\mathbf{A}^\top\right\|_\infty \|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_\infty + \left\|\mathbf{x}^* - \tilde{\mathbf{x}}_{kd+i}^*\right\|_\infty && \text{(Matrix Norm definition)} \\
&\leq \sqrt{m} 3^{-k} \alpha + 3^{-k} \kappa && \text{(From the Matrix Inequality 11)} \\
&\leq 2 \cdot 3^{-k} \kappa && \text{(From the Lemma statement, } \kappa \geq \sqrt{m}\alpha) \\
&\leq 3^{-(k-1)} \kappa
\end{aligned}$$

This sequence of inequalities demonstrates that $\mathbf{x} \in \mathcal{G}_{(k-1)d}^*$. Moreover, we know that $\mathcal{G}_{(k-1)d}^* \subseteq \mathcal{G}_{kd+i}^*$. Therefore, partitioning scheme $\mathcal{P}$ with $\kappa \geq \sqrt{m}\alpha$ will satisfy:

$$\mathcal{A}_{km+i}^* \subseteq \mathcal{P}_{kd+i}^* \quad \forall i \in [1 : m-1], \forall k \in \mathbb{N}_0$$

■

F.7. Proof of Lemma 16

Proof Since Lemma 15 is assumed to be applicable, consider a partitioning scheme $\mathcal{P}$ with $\kappa = \sqrt{m}\alpha$ and suppose $f_1^* = \sup_{\mathbf{x}\in\kappa\mathbb{H}_1^d} f(\mathbf{x})$. Let $\mathcal{P}$ satisfies theorem 12, then, there exist constants ν and $0 < \rho < 1$ such that

$$\sup_{\mathbf{x}\in\mathcal{P}_h^*} (f_1^* - f(\mathbf{x})) \leq \nu\rho^h. \quad \forall h \in \mathbb{N}_0 \tag{18}$$

Since $\kappa \geq \sqrt{m}\alpha$, we can incorporate Lemma 15, which gives,

$$\mathcal{A}_{km+i}^* \subseteq \mathcal{P}_{kd+i}^* \quad \forall i \in [1 : m-1], \forall k \in \mathbb{N}_0$$

Therefore, we get

$$\sup_{\mathbf{x}\in\mathcal{A}_{km+i}^*} (f^* - f(\mathbf{x})) \leq \sup_{\mathbf{x}\in\mathcal{P}_{kd+i}^*} (f^* - f(\mathbf{x})) \quad \forall i \in [1 : m-1], \forall k \in \mathbb{N}_0 \tag{19}$$

Combining eqs. (18) and (19), and using the fact $f^* \leq f_1^*$, we have

$$\sup_{\mathbf{x}\in\mathcal{A}_{km+i}^*} (f^* - f(\mathbf{x})) \leq \sup_{\mathbf{x}\in\mathcal{P}_{kd+i}^*} (f^* - f(\mathbf{x})) \leq \nu\rho^{kd+i} \leq \nu(\rho^{\frac{kd+i}{km+i}})^{km+i}$$

Therefore, we have

$$\sup_{\mathbf{x}\in\mathcal{A}_{km+i}^*} (f^* - f(\mathbf{x})) \leq \nu(\rho^{\frac{kd+i}{km+i}})^{km+i} \quad \forall i \in [1 : m-1], \forall k \in \mathbb{N}_0 \tag{20}$$

Ignoring first m heights and by choosing, $\nu_{\mathcal{A}} = \nu, \rho_{\mathcal{A}} = \rho^{\beta}$. Where

$$\beta = \min\left\{ \frac{kd+i}{km+i} \,\middle|\, i \in [1 : m-1],\ k \in \mathbb{N} \right\}$$

Now, we show $\beta = \frac{d+m-1}{2m-1}$.

To prove the above statement, we consider the sequence for $k = 1$:

$$t_i = \frac{d+i}{m+i}, \quad \forall i \in [1, m-1]$$

First, we show that this sequence is decreasing.

For any i in the range $[2, m-2]$, consider the difference between consecutive terms:

$$t_{i+1} - t_i = \frac{d+(i+1)}{m+(i+1)} - \frac{d+i}{m+i} = \frac{m-d}{(m+i)(m+i+1)}$$

Since $m > d$, we have $m - d > 0$, but the denominator $(m+i)(m+i+1)$ is positive. Thus, $t_{i+1} < t_i$, which says that sequence t_i is decreasing and the minimum value of t_i occurs at $i = m-1$:

$$t_{m-1} = \frac{d+(m-1)}{m+(m-1)} = \frac{d+m-1}{2m-1}$$

We now consider the general form for any $k \in \mathbb{N}$:

$$\beta = \min\left\{ \frac{kd+i}{km+i} \,\middle|\, i \in [1 : m-1] \right\}$$

With the observation that

$$\frac{(k+1)d+m-1}{(k+1)m+m-1} \geq \frac{kd+m-1}{km+m-1}$$

we get $\beta = (d+m-1)/(2m-1)$

With this choice of β we have

$$\sup_{\mathbf{x}\in\mathcal{A}^*_{km+i}} (f^* - f(\mathbf{x})) \leq \nu(\rho^{\frac{kd+i}{km+i}})^{km+i} \leq \nu_{\mathcal{A}}\rho_{\mathcal{A}}^{\beta} \quad \forall i \in [1:m-1], \forall k \in \mathbb{N} \tag{21}$$

Now, we choose slightly larger $\nu_{\mathcal{A}}$ to satisfy the first $m-1$ heights.

Using Equation 20,

$$\sup_{\mathbf{x}\in\mathcal{A}^*_i} (f^* - f(\mathbf{x})) \leq \nu\rho^i = \nu\rho^{i-\beta i}\rho^{\beta i} \quad \forall i \in [1:m-1] \tag{22}$$

$$\leq \nu\rho^{(1-\beta)(m-1)}\rho^{\beta i} \quad \forall i \in [1:m-1] \tag{23}$$

Therefore, $\nu_{\mathcal{A}} = \nu\rho^{(1-\beta)(m-1)}$ and $\rho_{\mathcal{A}} = \rho^{\beta}$ will satisfy the Assumption 12 for the $\mathcal{A}$ partitioning scheme.

■

F.8. Proof of Lemma 17

Proof Consider partitioning schemes $\mathcal{P}$ ($\kappa = 1$) and $\mathcal{G}$ ($\kappa = \kappa_1 = \sqrt{m}\alpha$) with star cells: $\mathcal{P}^*_h = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^d, \left\|\frac{\mathbf{x}-\mathbf{x}^*_h}{\mathbf{c}}\right\|_\infty \leq 1\}, \mathcal{G}^*_h = \{\mathbf{x} : \mathbf{x} \in \mathbb{R}^d, \left\|\frac{\mathbf{x}-\tilde{\mathbf{x}}^*_h}{\mathbf{c}}\right\|_\infty \leq \kappa\}$ where $\mathbf{c} = [3^{-\lfloor\frac{h+d-i}{d}\rfloor}]_{i=1}^d$.

Let $\tilde{h}_1 = d\lceil\log_3 \kappa_1\rceil$. We have $\mathcal{G}^*_{h+\tilde{h}_1} \subseteq \mathcal{P}^*_h$, implying:

$$\sup_{\mathbf{x}\in\mathcal{G}^*_{\tilde{h}_1+h}} (f^* - f(\mathbf{x})) \leq \sup_{\mathbf{x}\in\mathcal{P}^*_h} (f^* - f(\mathbf{x})) \quad \forall h \in \mathbb{N}_0 \tag{24}$$

By lemma assumption $\mathcal{P}$ satisfies Assumption 12, with parameters (ν, ρ). Therefore we have

$$\sup_{\mathbf{x}\in\mathcal{P}^*_h} (f^* - f(\mathbf{x})) \leq \nu\rho^h \quad \forall h \in \mathbb{N}_0$$

Therefore,

$$\sup_{\mathbf{x}\in\mathcal{G}^*_{\tilde{h}_1+h}} (f^* - f(\mathbf{x})) \leq \nu\rho^h = \nu\rho^{-\tilde{h}}\rho^{\tilde{h}+h} \quad \forall h \in \mathbb{N}_0 \tag{25}$$

For depths $h \in [1:\tilde{h}_1 - 1]$:

$$\sup_{\mathbf{x}\in\mathcal{G}^*_h} (f^* - f(\mathbf{x})) \leq f^* - \inf_{\mathbf{x}\in\kappa_1\mathbb{H}_1^d} f(\mathbf{x})$$

$$\leq (f^* - \inf_{\mathbf{x}\in\kappa_1\mathbb{H}_1^d} f(\mathbf{x}))\rho^{-\tilde{h}_1}\rho^h$$

Using inequality 25 and the above inequality, we conclude that $\mathcal{G}$ satisfies Assumption 12 with parameters $(\rho, \max\{\nu, f^* - \inf_{\mathbf{x}\in\kappa_1\mathbb{H}_1^d} f(\mathbf{x})\}\rho_1^{-\tilde{h}_1})$.

By Lemma 16 and$\kappa \geq \sqrt{m}\alpha$, we conclude:

$$\rho_{\mathcal{A}} = \rho^{\beta}, \nu_{\mathcal{A}} = \max\{\nu, f^* - \inf_{\mathbf{x}\in\kappa_1\mathbb{H}_1^d} f(\mathbf{x})\}\rho^{(1-\beta)(m-1)-\tilde{h}_1}$$

■

F.9. Proof of Lemma 19

Proof

Let $B^\circ(\mathbf{x}, r)$ be the open ball corresponding to the closed ball $B(\mathbf{x}, r)$. We will first demonstrate that $B^\circ(\mathbf{c}_i, \kappa_2) \cap B^\circ(\mathbf{c}_j, \kappa_2) = \emptyset \quad \forall i \neq j$. Suppose there exists a point $\mathbf{y} \in B^\circ(\mathbf{c}_i, \kappa_2) \cap B^\circ(\mathbf{c}_j, \kappa_2)$, then:

$$\|\mathbf{c}_i - \mathbf{c}_j\|_\infty \leq \|\mathbf{c}_i - \mathbf{y}\|_\infty + \|\mathbf{c}_j - \mathbf{y}\|_\infty < 2\kappa_2.$$

However, for lattice points, we have $\|\mathbf{c}_i - \mathbf{c}_j\|_\infty \geq 2\kappa_2$, contradicting the above inequality.

Next, we show that

$$B(\mathbf{c}_i, \kappa_2) \subseteq (\kappa_1 + 2\kappa_2)\mathbb{H}_1^m \quad \forall \mathbf{c}_i \in \mathcal{C}. \tag{26}$$

For any $\mathbf{y} \in B(\mathbf{c}_i, \kappa_2)$ we have that:

$$\begin{aligned}
\|\mathbf{y}\|_\infty &\leq \|\mathbf{y} - \mathbf{c}_i\|_\infty + \|\mathbf{c}_i\|_\infty \\
&\leq \kappa_2 + \|\mathbf{c}_i\|_\infty \quad \text{(since } \mathbf{y} \in B(\mathbf{x}_i, \kappa_2)\text{)} \\
&= \kappa_2 + \|\mathbf{c}_i - \mathbf{z} + \mathbf{z}\|_\infty \qquad \text{(for some } \mathbf{z} \in B(\mathbf{0}, \kappa_1) \cap B(\mathbf{c}_i, \kappa_2) \neq \emptyset\text{)} \\
&\leq \kappa_2 + \|\mathbf{c}_i - \mathbf{z}\|_\infty + \|\mathbf{z}\|_\infty \\
&\leq \kappa_2 + \kappa_2 + \kappa_1 \\
&= \kappa_1 + 2\kappa_2.
\end{aligned}$$

Therefore, $\mathbf{y} \in (\kappa_1 + 2\kappa_2)\mathbb{H}_1^m$, which proves (26). Let N be the number of balls $B(\mathbf{c}_i, \kappa_2)$. Since $B^\circ(\mathbf{x}_i, \kappa_2)$ are disjoint and all these balls are contained in $(\kappa_1 + 2\kappa_2)\mathbb{H}_1^m$, we have:

$$\begin{aligned}
N \cdot \text{Vol}(B^\circ(\mathbf{x}_i, \kappa_2)) &\leq \text{Vol}((\kappa_1 + 2\kappa_2)\mathbb{H}_1^m) \\
N \cdot (2\kappa_2)^m &\leq (\kappa_1 + 2\kappa_2)^m,
\end{aligned}$$

giving that $N \leq \left(2 + \frac{\kappa_1}{\kappa_2}\right)^m$. For the lower bound to N, since, $B(\mathbf{0}, \kappa_1) \subseteq \bigcup_i B(\mathbf{c}_i, \kappa_2)$, we have that

$$\begin{aligned}
\text{Vol}(B(\mathbf{0}, \kappa_1)) = (2\kappa_1)^m &\leq \text{Vol}(\bigcup_i B(\mathbf{c}_i, \kappa_2)) \\
&\leq \sum_{i=1}^N \text{Vol}(B(\mathbf{c}_i, \kappa_2)) = N\text{Vol}(B(\mathbf{c}_0, \kappa_2)) = N(2\kappa_2)^m.
\end{aligned}$$

Hence $N \geq \left(\frac{\kappa_1}{\kappa_2}\right)^m$. ■

F.10. Proof of Lemma 20

Proof

For the partitioning scheme $\mathcal{A}$, at a specific depth h, each cell can be indexed by i, i.e, $\mathcal{A}(h,i) \quad 1 \leq i \leq 3^h$, where, 3^h represents the total number of cells at depth h, and i is the index of a specific cell within that depth. Similarly, For the partitioning scheme $\mathcal{P}$, at a depth h, each cell can be indexed by j, i.e, $\mathcal{P}(h,j) \quad 1 \leq j \leq 3^h$.

Definition of the $POpt$ Relation Consider the following definition of the $POpt$ relation:

$$POpt = \{(\mathcal{A}(h,i), \mathcal{P}(h,j)) : \mathcal{A}(h,i) \in \mathcal{N}_{\mathcal{A}}(3\nu\rho^h),\ \mathcal{P}(h,j) \in \mathcal{N}_{\mathcal{P}}(3\nu\rho^h), \\ \mathcal{T}(h,i) \cap \mathbf{A}\mathcal{P}(h,j) \neq \emptyset\}$$

Here, $\mathcal{A}(h,i)$ represents a cell in the partition $\mathcal{A}$ at depth h indexed by i, and $\mathcal{T}(h,i)$ is equivalent partitioning scheme of $\mathcal{A}$. $\mathcal{P}(h,j)$ represents a cell in the partition $\mathcal{P}$ at depth h indexed by j. The sets $\mathcal{N}_{\mathcal{A}}(3\nu\rho^h)$ and $\mathcal{N}_{\mathcal{P}}(3\nu\rho^h)$ denote near-optimal cells in the respective partitions.

For any h, i, consider the cell $\mathcal{A}(h,i) \in \mathcal{N}_{\mathcal{A}}(3\nu\rho^h)$. Let l be the lower bound on the number of elements in $POpt$ that are of the form $(\mathcal{A}(h,i), \cdot)$. Then, we have: $|POpt| \geq |\mathcal{N}_{\mathcal{A}}(3\nu\rho^h)| \cdot l$. Similarly, for any h, i, consider the cell $\mathcal{P}(h,j) \in \mathcal{N}_{\mathcal{P}}(3\nu\rho^h)$. Let u be the upper bound on the number of elements in $POpt$ that are of the form $(\cdot, \mathcal{P}(h,j))$. Then, we have: $|\mathcal{N}_{\mathcal{P}}(3\nu\rho^h)| \cdot u \geq |POpt|$. Combining these two inequalities, we get:

$$|\mathcal{N}_{\mathcal{A}}(3\nu\rho^h)| \leq |\mathcal{N}_{\mathcal{P}}(3\nu\rho^h)| \cdot \frac{u}{l} \tag{27}$$

This inequality provides a relationship between the near-optimal cells in the two partitions, taking into account the bounds on the number of elements in the $POpt$ relation.

We define the following quantity to proceed with the proof.

$$\forall S \subseteq \mathbb{R}^d, \quad \text{Proj}_{\mathbf{A}_\perp}(S) = \{\boldsymbol{\beta} \in \mathbb{R}^{d-m} : \boldsymbol{\beta} = \mathbf{A}_\perp \mathbf{x}, \mathbf{x} \in S\} \tag{28}$$

Estimating Upper Bound u Consider a near-optimal cell $\mathcal{P}(h,j) \in \mathcal{N}_{\mathcal{P}}(3\nu\rho^h)$. Then the region $\mathbf{A}\mathcal{P}(h,j)$ is near optimal. To get the upper bound, we need to count how many cells of $\mathcal{T}(h,i)$ can fit into $\mathbf{A}\mathcal{P}(h,j)$ region. Using Proposition 14, the cell $\mathcal{T}(h,i)$ and $\mathcal{P}(h,j)$ can be written as

$$\mathcal{T}(h,i) = \{\boldsymbol{\alpha} \in \mathbb{R}^m, \left\| \frac{\boldsymbol{\alpha} - \boldsymbol{\alpha}_{h,i}}{\mathbf{s}} \right\|_\infty \leq \alpha \quad \text{with} \quad \mathbf{s} = [3^{-\lfloor \frac{h+m-j}{m} \rfloor}]_{j=1}^m\}$$

$$\mathcal{P}(h,i) = \{\mathbf{x} \in \mathbb{R}^d, \left\| \frac{\mathbf{x} - \mathbf{x}_{h,i}}{\mathbf{c}} \right\|_\infty \leq \kappa \quad \text{with} \quad \mathbf{c} = [3^{-\lfloor \frac{h+d-j}{d} \rfloor}]_{j=1}^d\}$$

For all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{P}(h,i)$, consider $\|\mathbf{A}\mathbf{x}_1 - \mathbf{A}\mathbf{x}_2\|_\infty$, we have:

$$\begin{aligned}
&= \|\mathbf{A}\mathbf{x}_1 - \mathbf{A}\mathbf{x}_2\|_\infty \\
&\leq \|\mathbf{A}\|_\infty \|\mathbf{x}_1 - \mathbf{x}_2\|_\infty && \text{(from the Matrix operator norm definition)} \\
&\leq \sqrt{m} \|\mathbf{x}_1 - \mathbf{x}_2\|_\infty && \text{(From Matrix Inequality 11 and } \|\mathbf{A}\|_2 = 1) \\
&= \sqrt{m} \|\mathbf{x}_1 - \mathbf{x}_2 + \mathbf{x}_{h,i} - \mathbf{x}_{h,i}\|_\infty \\
&\leq \sqrt{m} \|\mathbf{x}_1 - \mathbf{x}_{h,i}\|_\infty + \sqrt{m} \|\mathbf{x}_1 - \mathbf{x}_{h,i}\|_\infty \\
&\leq 2\sqrt{m} 3^{-k} && \text{(Let } h = kd + i)
\end{aligned}$$

Now, we estimate the side lengths for the $\mathcal{T}(h,i)$ cell, by computing $\mathbf{s}$

$$\frac{h+m-j}{m} = \frac{kd+m+i-j}{m} \leq \frac{kd+m+d}{m} \leq (k+1)\frac{d}{m}+1$$

Hence,

$$3^{-\left\lfloor \frac{h+m-j}{m} \right\rfloor} \geq 3^{-(1+(k+1)\frac{d}{m})}$$

Now we will invoke Lemma 19, to get the upper bound u Thus,

$$\begin{aligned} u &= \left(2+\frac{2\sqrt{m}3^{-k}}{\alpha 3^{-1}3^{-(k+1)\frac{d}{m}}}\right)^m = \left(2+\frac{6\sqrt{m}3^{d/m}3^{k(\frac{d-m}{m})}}{\alpha}\right)^m \\ &\leq \left(2+6\sqrt{m}3^{d/m}3^{k(\frac{d-m}{m})}\right)^m \\ &\leq (12\sqrt{m})^m 3^d 3^{k(d-m)} \end{aligned}$$

Estimating Lower Bound l The following set containment relation is used in the lower bound estimation. Denote, P, P' to be the domains $[-1,1]^d, [-\kappa,\kappa]^d$ respectively. Consider the matrix $\mathbf{Q} = \begin{bmatrix}\mathbf{A}^\top & \mathbf{A}_\perp^\top\end{bmatrix}$, where the columns of $\mathbf{Q}$ contain an orthonormal basis for $\mathbb{R}^d$. Given this matrix, we define the below quantities

$$\boldsymbol{\alpha}_{\max} \triangleq \begin{bmatrix}\max_{1\leq j\leq 2^d} \mathbf{q}_1^\top \mathbf{c}_j & \max_{1\leq j\leq 2^d} \mathbf{q}_2^\top \mathbf{c}_j & \cdots & \max_{1\leq j\leq 2^d} \mathbf{q}_m^\top \mathbf{c}_j\end{bmatrix}^\top$$

$$\boldsymbol{\alpha}'_{\max} \triangleq \begin{bmatrix}\max_{1\leq j\leq 2^d} \mathbf{q}_{m+1}^\top \mathbf{c}_j & \max_{1\leq j\leq 2^d} \mathbf{q}_{m+2}^\top \mathbf{c}_j & \cdots & \max_{1\leq j\leq 2^d} \mathbf{q}_d^\top \mathbf{c}_j\end{bmatrix}^\top$$

We introduce another domain P_{rot} to be the smallest rotated hyper-rectangle aligned with the columns of $\mathbf{Q}$ such that $P \subseteq P_{\text{rot}}$. We choose κ for the P' domain to ensure $P_{\text{rot}} \subseteq P'$.

Define the set of corners of P_{rot} as $\mathcal{C}_{\text{rot}}$ and we choose set of corners to be

$$\mathcal{C}_{\text{rot}} = \left\{\mathbf{A}^\top(\boldsymbol{\alpha}_{\max}\odot \mathbf{s}_1) + \mathbf{A}_\perp^\top(\boldsymbol{\alpha}'_{\max}\odot \mathbf{s}_2) : \mathbf{s}_1 \in \{-1,1\}^m, \mathbf{s}_2 \in \{-1,1\}^{d-m}\right\} \tag{29}$$

Since P' is an axis-aligned domain, $\kappa = \max_{\mathbf{c}_i\in\mathcal{C}_{\text{rot}}} \|\mathbf{c}_i\|_\infty$ will ensure $\mathcal{P}_{\text{rot}} \subseteq \mathcal{P}'(0,0)$. Bounding the κ:

$$\begin{aligned} \kappa &= \max_{\mathbf{c}\in\mathcal{C}_{\text{rot}}} \|\mathbf{c}\|_\infty \\ &= \max_{\mathbf{s}_1\in\{-1,1\}^m, \mathbf{s}_2\in\{-1,1\}^{d-m}} \left\|\mathbf{A}^\top(\boldsymbol{\alpha}_{\max}\odot \mathbf{s}_1) + \mathbf{A}_\perp^\top(\boldsymbol{\alpha}'_{\max}\odot \mathbf{s}_2)\right\|_\infty \\ &\leq \max_{\mathbf{s}_1\in\{-1,1\}^m, \mathbf{s}_2\in\{-1,1\}^{d-m}} \left\|\mathbf{A}^\top(\boldsymbol{\alpha}_{\max}\odot \mathbf{s}_1) + \mathbf{A}_\perp^\top(\boldsymbol{\alpha}'_{\max}\odot \mathbf{s}_2)\right\|_2 \\ &= \max_{\mathbf{s}_1\in\{-1,1\}^m, \mathbf{s}_2\in\{-1,1\}^{d-m}} \left\|\begin{bmatrix}\mathbf{A}^\top & \mathbf{A}_\perp^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}_{\max}\odot \mathbf{s}_1 \\ \boldsymbol{\alpha}'_{\max}\odot \mathbf{s}_2\end{bmatrix}\right\|_2 \\ &= \max_{\mathbf{s}_1\in\{-1,1\}^m, \mathbf{s}_2\in\{-1,1\}^{d-m}} \left\|\begin{bmatrix}\boldsymbol{\alpha}_{\max}\odot \mathbf{s}_1 \\ \boldsymbol{\alpha}'_{\max}\odot \mathbf{s}_2\end{bmatrix}\right\|_2 \quad \text{(Orthonormal matrix does not change the length)} \\ &= \left\|\begin{bmatrix}\boldsymbol{\alpha}_{\max} \\ \boldsymbol{\alpha}'_{\max}\end{bmatrix}\right\|_2 \leq d \qquad (\boldsymbol{\alpha}_{\max_j} = \max_{1\leq i\leq 2^d} |\mathbf{a}_j\mathbf{c}_i| \leq \|\mathbf{a}_j\|_2\|\mathbf{c}_i\|_2 = \sqrt{d}.) \end{aligned}$$

Hence, the domain P' with $\kappa \geq d$ guarantees $P_{\text{rot}} \subseteq P'$ and we have $P \subseteq P_{\text{rot}} \subseteq P'$. Moreover, $\text{Vol}(\text{Proj}_{\mathbf{A}_\perp}(P_{\text{rot}})) = \prod_{i=1}^{d-m} 2\boldsymbol{\alpha}'_{\max}$. Hence,

$$\text{Vol}((\text{Proj}_{\mathbf{A}_\perp}(P'_{\text{rot}}))) \geq \prod_{i=1}^{d-m} 2\boldsymbol{\alpha}'_{\max} \tag{30}$$

Now consider a cell $\mathcal{A}(h,i) \in \mathcal{N}_{\mathcal{A}}(3\nu\rho^h)$. i.e, $\exists \mathbf{x} \in \mathcal{A}(h,i)$ for which $f(\mathbf{x}) \geq f(\mathbf{x}^*) - 3\nu\rho^h$. We find the number of pairs in $POpt$ which contain $\mathcal{A}(h,i)$ by counting all cells in $\mathcal{P}$ whose projection onto A contains $\mathbf{x}$. This number n_2 can be lower bounded by dividing the volume of the $d-m$ dimensional orthogonal space associated with $\mathbf{x}$ by the volume of the projection onto the orthogonal space of a cell in $\mathcal{P}(h,0)$. The number n_2 can be lower bounded as

$$\begin{aligned}
n_2 &\geq \frac{\text{Vol}(\text{Proj}_{\mathbf{A}_\perp}(\mathcal{P}(0,0)))}{\text{Vol}(\text{Proj}_{\mathbf{A}_\perp}(\mathcal{P}(h,0)))} \\
&\geq \frac{\prod_{i=1}^{d-m}(1/d)2(\boldsymbol{\alpha}'_{\max})_i}{\prod_{i=m+1}^{d} 2\max_{1\leq j\leq 2^d} \mathbf{q}_i^\top \mathbf{x}_j} && \text{(Using 30 on } (1/d) \text{ scaled domain)} \\
&\geq \frac{\prod_{i=m+1}^{d}(1/d)2\max_{1\leq j\leq 2^d} \mathbf{q}_i^\top \mathbf{c}_j}{\prod_{i=m+1}^{d} 2\max_{1\leq j\leq 2^d} \mathbf{q}_i^\top \mathbf{x}_j} \\
&\geq \frac{(1/d)^{d-m}\prod_{i=m+1}^{d}\max_{1\leq j\leq 2^d} \mathbf{q}_i^\top \mathbf{c}_j}{\prod_{i=m+1}^{d} 3^{-k}\max_{1\leq j\leq 2^d} \mathbf{q}_i^\top \mathbf{c}_j} \\
&\geq \frac{1}{3^{-k(d-m)}} \geq (1/d)^{d-m}3^{k(d-m)}.
\end{aligned}$$

Hence, the lower bound $l = (1/d)^{d-m}3^{k(d-m)}$

Thus, using Equation 27, we conclude that

$$|\mathcal{N}_{\mathcal{A}}(3\nu\rho^h)| \leq |\mathcal{N}_{\mathcal{T}}(3\nu\rho^h)| \cdot 3^d d^{d-m}(12\sqrt{m})^m \tag{31}$$

■

F.11. Proof of Lemma 21

Proof

To start, we recall the relationship between the parameters of the two partitioning schemes as established in Lemma 17. Specifically, we have:

$$\rho_{\mathcal{A}} = \rho^\beta, \nu_{\mathcal{A}} = \max\{\nu, f^* - \inf_{\mathbf{x}\in\kappa_1\mathbb{H}_1^d} f(\mathbf{x})\}\rho^{(1-\beta)(m-1)-\tilde{h}_1}$$

For brevity, we denote $\nu' = \max\{\nu, f^* - \inf_{\mathbf{x}\in\kappa_1\mathbb{H}_1^d} f(\mathbf{x})\}$.

From the definiton of near-optimality dimension for the $\mathcal{P}$ partioning scheme, we have:

$$\mathcal{N}_{\mathcal{P}}(3\nu\rho^h) \leq C\rho^{-\eta_{\mathcal{P}}\, h} \tag{32}$$

Now, consider $3\nu_{\mathcal{A}}\rho_{\mathcal{A}}^{h}$:

$$
\begin{aligned}
&= 3\nu'\rho^{\beta h+(1-\beta)(m-1)-\tilde{h}_1} \\
&= 3\nu\rho^{\beta h+\log_\rho(\nu'\rho^{(1-\beta)(m-1)-\tilde{h}_1})-\log_\rho(\nu)} \\
&\leq 3\nu\rho^{h+\log_\rho(\nu'\rho^{(1-\beta)(m-1)-\tilde{h}_1})-\log_\rho(\nu)} \\
&\leq 3\nu\rho^{h+\left\lfloor\log_\rho(\nu'\rho^{(1-\beta)(m-1)-\tilde{h}_1})-\log_\rho(\nu)\right\rfloor} \\
&= 3\nu\rho^{h-\tilde{h}_3} \qquad \text{(Denote } \tilde{h}_3 = -\left\lfloor\log_\rho(\nu'\rho^{(1-\beta)(m-1)-\tilde{h}_1}) - \log_\rho(\nu)\right\rfloor\text{)}
\end{aligned}
$$

Now, consider

$$
\begin{aligned}
\mathcal{N}_{\mathcal{A}}(3\nu_{\mathcal{A}}/\rho_{\mathcal{A}}^{-h}) &\leq \mathcal{N}_{\mathcal{A}}(3\nu\rho^{h-\tilde{h}_3}) && (3\nu_{\mathcal{A}}\rho_{\mathcal{A}}^{h} \leq 3\nu\rho^{h-\tilde{h}_3}) \\
&\leq 3^d d^{d-m}(12\sqrt{m})^m \mathcal{N}_{\mathcal{P}}(3\nu\rho^{h-\tilde{h}_3}) && \text{(Using Lemma 20)} \\
&\leq 3^d d^{d-m}(12\sqrt{m})^m C\rho^{-\eta_{\mathcal{P}}(h-\tilde{h}_3)} && \text{(Using Inequality 32)} \\
&= 3^d d^{d-m}(12\sqrt{m})^m C\rho^{\eta_{\mathcal{P}}\tilde{h}_3}\rho^{-\eta_{\mathcal{P}}h}
\end{aligned}
$$

Therefore, we have:

$$
\mathcal{N}_{\mathcal{A}}(3\nu_{\mathcal{A}}/\rho_{\mathcal{A}}^{-h}) \leq 3^d d^{d-m}(12\sqrt{m})^m C\rho^{\eta_{\mathcal{P}}\tilde{h}_3}\rho^{-\eta_{\mathcal{P}}h} \quad \forall h \geq \tilde{h}_3
$$

For heights $h \in [0 : \tilde{h}_3 - 1]$, since the right-hand side quantity is monotonically increasing, we can use the value of the right-hand side at depth $h = \tilde{h}_3$, which is $3^d d^{d-m}(12\sqrt{m})^m C$.

Therefore, we have

$$
\mathcal{N}_{\mathcal{A}}(3\nu_{\mathcal{A}}/\rho_{\mathcal{A}}^{-h}) \leq 3^d d^{d-m}(12\sqrt{m})^m C\rho^{-\eta_{\mathcal{P}}\tilde{h}_3}\rho^{-\eta_{\mathcal{P}}h} = 3^d d^{d-m}(12\sqrt{m})^m C\rho^{-\eta_{\mathcal{P}}\tilde{h}_3}\rho_{\mathcal{A}}^{-\frac{\eta_{\mathcal{P}}}{\beta}h}
$$

This implies that $\eta_{\mathcal{A}} \leq \eta_{\mathcal{P}}/\beta$, and the constant $C_{\mathcal{A}}$ is given by $3^d d^{d-m}(12\sqrt{m})^m C\rho^{-\eta_{\mathcal{P}}\tilde{h}_3}$. ■

For some of the proofs, we use the following equivalence of partitioning schemes.

F.12. Equivalence of partitioning schemes

Consider the relationship between partitioning schemes $\mathcal{A}$ and $\mathcal{T}$ as defined in Definition 3. For every $\boldsymbol{\alpha} \in \mathcal{T}_{h,i}$, there exists a unique $\mathbf{x} \in \mathcal{A}_{h,i}$ such that $\mathbf{x} = \mathbf{A}^\top\boldsymbol{\alpha}$. This establishes an equivalence:

$$
\forall \mathbf{x} \in \mathcal{A}_{h,i}, \quad f(\mathbf{x}) = f(\mathbf{A}^\top\boldsymbol{\alpha}) = g(\mathbf{A}\mathbf{A}^\top\boldsymbol{\alpha}) = g(\boldsymbol{\alpha}) \tag{33}
$$

where f is optimized on $\mathcal{A}_{h,i}$ and g on $\mathcal{T}_{h,i}$. Thus, optimizing f over $\mathcal{A}$ is equivalent to optimizing g over $\mathcal{T}$.

Similarly, for $\hat{\mathcal{A}}$ and $\hat{\mathcal{T}}$, we have: $\forall\boldsymbol{\alpha} \in \hat{\mathcal{T}}_{h,i}, \exists\mathbf{x} \in \hat{\mathcal{A}}_{h,i}$ such that $\mathbf{x} = \hat{\mathbf{A}}^\top\boldsymbol{\alpha}$. This leads to:

$$
\forall \mathbf{x} \in \hat{\mathcal{A}}_{h,i}, \quad f(\mathbf{x}) = f(\hat{\mathbf{A}}^\top\boldsymbol{\alpha}) = g(\mathbf{A}\hat{\mathbf{A}}^\top\boldsymbol{\alpha}) \stackrel{\text{def}}{=} \hat{g}(\boldsymbol{\alpha}) \tag{34}
$$

where g is defined on $\mathcal{T}_{h,i}$ and $\hat{g}$ on $\hat{\mathcal{T}}_{h,i}$. Therefore, optimizing f over $\hat{\mathcal{A}}$ is equivalent to optimizing $\hat{g}$ over $\hat{\mathcal{T}}$.

F.13. Proof of Lemma 28

Proof

From the Notation section A, we have $\mathbf{W}_\perp = \mathbf{W} - \mathbf{W}\mathbf{A}\mathbf{A}^\top$. Next, consider the singular value decomposition (SVD) of $\mathbf{W}$, given by $\mathbf{W} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{p\times p}$ and $\mathbf{V} \in \mathbb{R}^{d\times d}$ are orthogonal matrices, satisfying $\mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top\mathbf{U} = \mathbf{I}_p, \mathbf{V}\mathbf{V}^\top = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_d$ and $\mathbf{S} \in \mathbb{R}^{p\times d}$ is a diagonal matrix with diagonal elements.

We have this relation $\|\mathbf{W}\|_2 = \sigma_1(\mathbf{W}) \geq \cdots \geq \sigma_r(\mathbf{W}) \geq 0, \quad r = \min\{p, d\}$. From the Lemma assumption, we have $\operatorname{rank}(\mathbf{W}) = r \geq m$. Thus, we collect the respective leading m columns of $\mathbf{U}$ and $\mathbf{V}$, denoted as $\mathbf{U}_1 \in \mathbb{R}^{p\times m}$ and $\mathbf{V}_1 \in \mathbb{R}^{d\times m}$, respectively. The remaining columns are denote by $\mathbf{U}_2 \in \mathbb{R}^{p\times p-m}$ and $\mathbf{V}_2 \in \mathbb{R}^{d\times d-m}$. We have:

$$\mathbf{U}_1^\top\mathbf{U}_1 = \mathbf{I}_m \quad \text{and} \quad \mathbf{U}_1^\top\mathbf{U}_2 = \mathbf{0}_{m\times p-m}. \tag{35}$$

Then the SVD of $\mathbf{W}$ can be written as: $\mathbf{W} = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^\top + \mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^\top$. and let $\hat{\mathbf{A}} = \mathbf{V}_1^\top$. Now, consider the implications of this decomposition for the proof. Consider,

$$\begin{aligned}
&= \left\|\mathbf{W}(\mathbf{I} - \mathbf{A}^\top\mathbf{A})\right\|_F \\
&\geq \left\|\mathbf{W}(\mathbf{I} - \mathbf{A}^\top\mathbf{A})\right\|_2 && \text{(From Matrix Norm Inequality 11)} \\
&= \left\|\mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^\top(\mathbf{I} - \mathbf{A}^\top\mathbf{A}) + \mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^\top(\mathbf{I} - \mathbf{A}^\top\mathbf{A})\right\|_2 && \text{(SVD of } \mathbf{W}\text{)} \\
&= \left\|\mathbf{U}_1^\top\right\|_2 \left\|\mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^\top(\mathbf{I} - \mathbf{A}^\top\mathbf{A}) + \mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^\top(\mathbf{I} - \mathbf{A}^\top\mathbf{A})\right\|_2 && (\left\|\mathbf{U}_1^\top\right\|_2 = 1) \\
&\geq \left\|\mathbf{U}_1^\top(\mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^\top(\mathbf{I} - \mathbf{A}^\top\mathbf{A}) + \mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^\top(\mathbf{I} - \mathbf{A}^\top\mathbf{A}))\right\|_2 && \text{(From Matrix Norm Inequality 12)} \\
&= \left\|\mathbf{S}_1\mathbf{V}_1^\top(\mathbf{I} - \mathbf{A}^\top\mathbf{A})\right\|_2 && \text{(Using Equation 35)} \\
&\geq \sigma_{min}(\mathbf{S}_1)\left\|\mathbf{V}_1^\top(\mathbf{I} - \mathbf{A}^\top\mathbf{A})\right\|_2 && \text{(From Matrix Norm Inequality 12)} \\
&= \sigma_m\left\|\hat{\mathbf{A}}(\mathbf{I} - \mathbf{A}^\top\mathbf{A})\right\|_2 \\
&= \sigma_m\left\|\hat{\mathbf{A}}\mathbf{A}_\perp^\top\mathbf{A}_\perp\right\|_2 && (\left[\mathbf{A}^\top, \mathbf{A}_\perp^\top\right] \text{ is a orthonormal matrix, thus } \mathbf{A}^\top\mathbf{A} + \mathbf{A}_\perp^\top\mathbf{A}_\perp = \mathbf{I}_d) \\
&= \sigma_m\left\|\hat{\mathbf{A}}\mathbf{A}_\perp^\top\mathbf{A}_\perp\right\|_2\left\|\mathbf{A}_\perp^\top\right\|_2 && (\left\|\mathbf{A}_\perp^\top\right\|_2 = 1) \\
&\geq \sigma_m\left\|\hat{\mathbf{A}}\mathbf{A}_\perp^\top\mathbf{A}_\perp\mathbf{A}_\perp^\top\right\|_2 && \text{(From Matrix Norm Inequality 12)} \\
&= \sigma_m\left\|\hat{\mathbf{A}}\mathbf{A}_\perp^\top\right\|_2 && (\mathbf{A}_\perp \text{ consists of orthonormal rows}) \\
&= \sigma_m\left\|\sin\Theta(\mathbf{A}, \hat{\mathbf{A}})\right\|_2 && \text{(From Definition 22)}
\end{aligned}$$

Since $\sigma_m > 0$, we have

$$\left\|\sin\Theta(\mathbf{A}, \hat{\mathbf{A}})\right\|_2 \leq \frac{\left\|\mathbf{W}(\mathbf{I} - \mathbf{A}^\top\mathbf{A})\right\|_2}{\sigma_m} \leq \frac{\left\|\mathbf{W}(\mathbf{I} - \mathbf{A}^\top\mathbf{A})\right\|_F}{\sigma_m}.$$

■

F.14. Proof of Lemma 24

Proof

Using the partitioning scheme equivalence from Section F.12, we can work with partitioning schemes $\mathcal{T}$ and $\hat{\mathcal{T}}$ in place of the $\mathcal{A}$ and $\hat{\mathcal{A}}$ partitioning schemes.

Let $\mathbf{z}_h, \hat{\mathbf{z}}_h$ be the representatives of $\mathcal{T}_h^*$ and $\hat{\mathcal{T}}_h^*$ cells respectively. Then

$$\mathcal{T}_h^* = \left\{ \mathbf{z} \in \mathbb{R}^m, \left\| \frac{\mathbf{z} - \mathbf{z}_h}{\mathbf{s}} \right\|_\infty \leq \alpha \quad \text{with} \quad \mathbf{s} = [3^{-\lfloor \frac{h+m-i}{m} \rfloor}]_{i=1}^m \right\}$$

and

$$\hat{\mathcal{T}}_h^* = \left\{ \mathbf{z} \in \mathbb{R}^m, \left\| \frac{\mathbf{z} - \hat{\mathbf{z}}_h}{\mathbf{s}} \right\|_\infty \leq \frac{\sqrt{m}\alpha}{\sqrt{1 - \text{dist}^2(\mathbf{A}, \hat{\mathbf{A}})}} \quad \text{with} \quad \mathbf{s} = [3^{-\lfloor \frac{h+m-i}{m} \rfloor}]_{i=1}^m \right\}$$

For brevity, denote $\kappa_2 = \frac{\sqrt{m}}{\sqrt{1-\text{dist}^2(\mathbf{A},\hat{\mathbf{A}})}}$. Let $\mathbf{z}^*$ denote a maximizer of the function g, i.e., $\mathbf{z}^* \in \arg\max_{\mathbf{z} \in \alpha \mathbb{H}_1^m} g(\mathbf{z})$. Given that $\hat{g}(\mathbf{z}) = g(\mathbf{A}\hat{\mathbf{A}}^\top \mathbf{z})$ and that $\mathbf{A}\hat{\mathbf{A}}^\top$ an invertible matrix, we can identify a corresponding maximizer $\mathbf{z}_{\hat{g}}^*$ for $\hat{g}$ such that: $\mathbf{z}_g^* = \mathbf{A}\hat{\mathbf{A}}^\top \mathbf{z}_{\hat{g}}^*$. Under this transformation, $\hat{g}(\mathbf{z}_{\hat{g}}^*)$ will be a maximizer of the function $\hat{g}$.

From lemma assumption, $\mathcal{T}$ satisfies Assumption 12, therefore we have:

$$\forall h \in \mathbb{N}_0, \sup_{\mathbf{z} \in \mathcal{T}_h^*} (g^* - g(\mathbf{z})) \leq \nu_{\mathcal{A}} \rho_{\mathcal{A}}^h \tag{36}$$

We aim to control

$$\forall h \in \mathbb{N}_0, \sup_{\mathbf{z} \in \hat{\mathcal{T}}_h^*} (g^* - \hat{g}(\mathbf{z})) = \sup_{\mathbf{z} \in \hat{\mathcal{T}}_h^*} (g^* - g(\mathbf{A}\hat{\mathbf{A}}^\top \mathbf{z})) \tag{37}$$

We choose some height h and represent height h as $h = km + i$. Lets denote $R_h = \{\mathbf{A}\hat{\mathbf{A}}^\top \mathbf{z} : \mathbf{z} \in \hat{\mathcal{T}}_h^*\}$. Now, we show that $R_{km} \subseteq \mathcal{T}_{km-mk'}^*$ where $k' = \left\lceil \log_3 \frac{2\sqrt{m}\kappa_2}{\kappa_2 - 1} \right\rceil$.

Consider a point $\mathbf{A}\hat{\mathbf{A}}^\top \mathbf{z}$ from the set R_{km}, where $\mathbf{z} \in \hat{\mathcal{T}}_{km}^*$. By definition, $\mathbf{z}$ satisfies the following condition:

$$\|\mathbf{z} - \hat{\mathbf{z}}_{km}\|_\infty \leq 3^{-k} \kappa_2 \alpha \tag{38}$$

Now, we show $\mathbf{A}\hat{\mathbf{A}}^\top \mathbf{z} \in \mathcal{T}_{km-mk'}^*$ We begin with the following inequality:

$$
\begin{aligned}
&= \left\|\mathbf{z}_{(k-k')m} - \mathbf{A}\hat{\mathbf{A}}^\top \mathbf{z}\right\|_\infty \\
&= \left\|\mathbf{z}_{(k-k')m} - \mathbf{z}^*_g + \mathbf{A}\hat{\mathbf{A}}^\top \mathbf{z}^*_{\hat{g}} - \mathbf{A}\hat{\mathbf{A}}^\top \mathbf{z}\right\|_\infty && \text{(Using } \mathbf{z}^*_g = \mathbf{A}\hat{\mathbf{A}}^\top \mathbf{z}^*_{\hat{g}}\text{)} \\
&\leq \left\|\mathbf{z}_{(k-k')m} - \mathbf{z}^*_g\right\|_\infty + \left\|\mathbf{A}\hat{\mathbf{A}}^\top (\mathbf{z}^*_{\hat{g}} - \mathbf{z})\right\|_\infty && \text{(Triangle Inequality of Norms)} \\
&\leq 3^{-(k-k')}\alpha + \left\|\mathbf{A}\hat{\mathbf{A}}^\top (\mathbf{z}^*_{\hat{g}} - \mathbf{z})\right\|_\infty && (\mathbf{z}^*_g \in \mathcal{T}^*_{km} \text{ cell. Thus } \|\mathbf{z}_{km} - \mathbf{z}^*\|_\infty \leq 3^{-k}\alpha) \\
&\leq 3^{-(k-k')}\alpha + \left\|\mathbf{A}\hat{\mathbf{A}}^\top\right\|_\infty \left\|\mathbf{z}^*_{\hat{g}} - \mathbf{z}\right\|_\infty && \text{(Operator Norm Definition)} \\
&\leq 3^{-(k-k')}\alpha + \sqrt{m}\left\|\mathbf{z}^*_{\hat{g}} - \mathbf{z}\right\|_\infty && \text{(From Matrix Inequality 11, 12 and } \|\mathbf{A}\|_2 = \left\|\hat{\mathbf{A}}\right\|_2 = 1) \\
&\leq 3^{-(k-k')}\alpha + \sqrt{m}\left\|-\hat{\mathbf{z}}_{km} + \mathbf{z}^*_{\hat{g}}\right\|_\infty + \sqrt{m}\left\|\hat{\mathbf{z}}_{km} - \mathbf{z}\right\|_\infty \\
&\leq 3^{-(k-k')}\alpha + \sqrt{m}3^{-k}2\kappa_2\alpha && \text{(using Inequality 38 and } \mathbf{z}^*_{\hat{g}} \in \hat{\mathcal{T}}^*{}_{km}\text{)} \\
&\leq 3^{-(k-k')}\alpha + 3^{-(k-k')}\alpha(\kappa_2 - 1) && \text{(Suppose } k' \text{ is chosen such that } 2\sqrt{m}\kappa_2 \leq (\kappa_2 - 1)3^{k'}\text{)} \\
&= \alpha\kappa_2 3^{-(k-k')}
\end{aligned}
$$

From the above inequality, we can conclude:

$$R_{km} \subseteq \mathcal{T}^*_{km - m\left\lceil \log_3 \frac{2\sqrt{m}\kappa_2}{\kappa_2 - 1}\right\rceil}$$

Using the above set containment, for any height $h = km + i$, we have

$$R_{km+i} \subseteq R_{km} \subseteq \mathcal{T}^*_{km - m\left\lceil \log_3 \frac{2\sqrt{m}\kappa_2}{\kappa_2 - 1}\right\rceil} \subseteq \mathcal{T}^*_{km - m\left\lceil \log_3 \frac{2\sqrt{m}\kappa_2}{\kappa_2 - 1}\right\rceil + i - m}$$

First and the last set containment are valid from the round robin paritioning scheme. Substituting, $h = km + i$, we get:$R_h \subseteq \mathcal{T}^*_{h - m\left\lceil \log_3 \frac{2\sqrt{m}\kappa_2}{\kappa_2 - 1}\right\rceil - m}$

Let us define: $\tilde{h}_2 = m\left\lceil \log_3 \frac{2\sqrt{m}\kappa_2}{\kappa_2 - 1}\right\rceil + m$. Using the Inequalities 36, 37 and the above set containment, we conclude:

$$\sup_{\mathbf{z} \in \hat{\mathcal{T}}^*_h} (g^* - \hat{g}(\mathbf{z})) \leq \nu_{\mathcal{A}} \rho_{\mathcal{A}}^{h - \tilde{h}_2} \quad \forall h \geq \tilde{h}_2 \tag{39}$$

For height $h \in [0 : \tilde{h}_2 - 1]$, we know that $\inf_{\mathbf{z} \in \hat{\mathcal{T}}^*_h} \hat{g}(\mathbf{z}) \geq \inf_{\mathbf{z} \in \hat{\mathcal{T}}^*_0} \hat{g}(\mathbf{z})$, which gives

$$
\begin{aligned}
\sup_{\mathbf{z} \in \hat{\mathcal{T}}^*_h} (g^* - \hat{g}(\mathbf{z})) &\leq g^* - \inf_{\mathbf{z} \in \hat{\mathcal{T}}^*_0} \hat{g}(\mathbf{z}) \\
&\leq (g^* - \inf_{\mathbf{z} \in \hat{\mathcal{T}}^*_0} \hat{g}(\mathbf{z}))\rho_{\mathcal{A}}^{-\tilde{h}_2}\rho_{\mathcal{A}}^{h}
\end{aligned}
$$

Using inequality 39 and the above inequality, we conclude that $\hat{\mathcal{T}}$ satisfies Assumption 12 with parameters

$$(\rho_{\mathcal{A}}, \max\{\nu_{\mathcal{A}}, g^* - \inf_{\mathbf{z} \in \kappa_2\alpha\mathbb{H}^m_1} g(\mathbf{A}\hat{\mathbf{A}}\mathbf{z})\}\rho_{\mathcal{A}}^{-\tilde{h}_2})$$

■

F.15. Proof of Lemms 23

Proof Consider $\mathbf{z}^* = (\mathbf{A}\hat{\mathbf{A}}^\top)^{-1}\mathbf{A}\mathbf{x}^*$, then

$$\begin{aligned} f(\hat{\mathbf{A}}^\top \mathbf{z}^*) &= g(\mathbf{A}\hat{\mathbf{A}}^\top \mathbf{z}^*) = g(\mathbf{A}\hat{\mathbf{A}}^\top(\mathbf{A}\hat{\mathbf{A}}^\top)^{-1}\mathbf{A}\mathbf{x}^*) \\ &= g(\mathbf{A}\mathbf{x}^*) = f(\mathbf{x}^*). \end{aligned}$$

Now, we show that

$$\begin{aligned} \|\mathbf{z}^*\|_\infty &= \left\|(\mathbf{A}\hat{\mathbf{A}}^\top)^{-1}\mathbf{A}\mathbf{x}^*\right\|_\infty \\ &\leq \left\|(\mathbf{A}\hat{\mathbf{A}}^\top)^{-1}\mathbf{A}\mathbf{x}^*\right\|_2 && \text{(Vector Norm Inequality)} \\ &\leq \left\|(\mathbf{A}\hat{\mathbf{A}}^\top)^{-1}\right\|_2 \|\mathbf{A}\mathbf{x}^*\|_2 && \text{(Operator Norm Definition)} \\ &= \frac{\left\|(\mathbf{A}\hat{\mathbf{A}}^\top)^{-1}\right\|_2 \sigma_{\min}(\mathbf{A}\hat{\mathbf{A}}^\top)}{\sigma_{\min}(\mathbf{A}\hat{\mathbf{A}}^\top)} \|\mathbf{A}\mathbf{x}^*\|_2 && \text{(Since, } \sigma_{\min}(\mathbf{A}\hat{\mathbf{A}}^\top) > 0) \\ &\leq \frac{\left\|(\mathbf{A}\hat{\mathbf{A}}^\top)^{-1}\mathbf{A}\hat{\mathbf{A}}^\top\right\|_2}{\sigma_{\min}(\mathbf{A}\hat{\mathbf{A}}^\top)} \|\mathbf{A}\mathbf{x}^*\|_2 && \text{(From Matrix Inequality 12)} \\ &\leq \frac{\|\mathbf{A}\mathbf{x}^*\|_2}{1 - \mathrm{dist}^2(\mathbf{A}, \hat{\mathbf{A}})} && \text{(From the Equation 8 and } \|\mathbf{I}\|_2 = 1) \\ &\leq \frac{\sqrt{m}\,\|\mathbf{A}\mathbf{x}^*\|_\infty}{1 - \mathrm{dist}^2(\mathbf{A}, \hat{\mathbf{A}})} && \text{(Vector Norm Inequality)} \\ &\leq \frac{\sqrt{m}\alpha}{1 - \mathrm{dist}^2(\mathbf{A}, \hat{\mathbf{A}})}. \end{aligned}$$

■

F.16. Proof of Proposition 11

Proof

Let $\mathbf{A} = [\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \ldots, \mathbf{w}_p] \in \mathbb{R}^{d\times p}$, where we assume without loss of generality that the vectors $\{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_p\}$ are linearly independent. If the vectors were dependent, we could consider only the independent vectors without changing the $\mathrm{Span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p\}$. The orthogonal projection matrix $\mathbf{P}$ onto $\mathrm{Span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p\}$ is given by $\mathbf{P} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$.

For all $\mathbf{x} \in \mathbb{R}^d$, consider $(\mathbf{w}_i^T\mathbf{P})\mathbf{x} = (\mathbf{P}^\top\mathbf{w}_i)^\top\mathbf{x} = (\mathbf{P}\mathbf{w}_i)^\top\mathbf{x} = \mathbf{w}_i^\top\mathbf{x}$, where the equalities hold due to the following: First, $\mathbf{P}^\top = \mathbf{P}$ because $\mathbf{P}$ is a projection matrix and thus symmetric. Second, $\mathbf{P}\mathbf{w}_i = \mathbf{w}_i$ since $\mathbf{w}_i$ is in the column span of $\mathbf{A}$, and $\mathbf{P}$ projects onto this span. Therefore, we can express $f(\mathbf{x})$ as

$$f(\mathbf{x}) = \sum_{i=1}^{p} v_i\sigma(\mathbf{w}_i^\top\mathbf{x} + b_i) = \sum_{i=1}^{p} v_i\sigma(\mathbf{w}_i^\top\mathbf{P}\mathbf{x} + b_i). \tag{40}$$

Since $(\mathbf{x} - \mathbf{x}') \perp \mathrm{Span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p\}, \mathbf{P}\mathbf{x} = \mathbf{P}\mathbf{x}'$ and therefore using eq. (40), $f(\mathbf{x}) = f(\mathbf{x}')$.
■

F.17. Proof of Lemma 25

Proof

Using the partitioning scheme equivalence from Section F.12, we can work with partitioning schemes $\mathcal{T}$ and $\hat{\mathcal{T}}$ in place of the $\mathcal{A}$ and $\hat{\mathcal{A}}$ partitioning schemes.

Consider an arbitrary cell $\mathcal{T}_{h,i}$ and $\hat{\mathcal{T}}_{h,i}$

$$\mathcal{T}_{h,i} = \left\{ \mathbf{z} : \mathbf{z} \in \mathbb{R}^m, \left\| \frac{\mathbf{z} - \mathbf{z}_{h,i}}{\mathbf{s}} \right\|_\infty \leq \alpha \quad \text{with} \quad \mathbf{s} = [3^{-\lfloor \frac{h+m-i}{m} \rfloor}]_{i=1}^m \right\} \tag{41}$$

and

$$\hat{\mathcal{T}}_{h,i} = \left\{ \mathbf{z} : \mathbf{z} \in \mathbb{R}^m, \left\| \frac{\mathbf{z} - \tilde{\mathbf{z}}_{h,i}}{\mathbf{s}} \right\|_\infty \leq \frac{\alpha\sqrt{m}}{\sqrt{1 - \mathrm{dist}^2(\mathbf{A}, \hat{\mathbf{A}})}} \quad \text{with} \quad \mathbf{s} = [3^{-\lfloor \frac{h+m-i}{m} \rfloor}]_{i=1}^m \right\} \tag{42}$$

We optimize $\hat{g}(\mathbf{z}) = g(\mathbf{A}\hat{\mathbf{A}}^\top \mathbf{z})$ on $\hat{\mathcal{T}}$ and $g(\mathbf{z})$ on $\mathcal{T}$. For a near-optimal cell $\mathcal{T}_{h,i} \in \mathcal{N}_{\mathcal{T}}(3\nu_{\mathcal{T}}\rho_{\mathcal{T}}^h)$, we have:

$$\sup_{\mathbf{z} \in \mathcal{T}_{h,i}} g(\mathbf{z}) = \sup_{\mathbf{z} \in (\mathbf{A}\hat{\mathbf{A}}^\top)^{-1}\mathcal{T}_{h,i}} g(\mathbf{A}\hat{\mathbf{A}}^\top \mathbf{z}) \geq g^* - 3\nu_{\mathcal{T}}\rho_{\mathcal{T}}^h \tag{43}$$

We aim to count the number of cells in $\hat{\mathcal{T}}$ that satisfy:

$$\sup_{\mathbf{z} \in \hat{\mathcal{T}}_{h,i}} g(\mathbf{A}\hat{\mathbf{A}}^\top \mathbf{z}) \geq g^* - 3\nu_{\mathcal{T}}\rho_{\mathcal{T}}^h \tag{44}$$

Using Relations 43, 44, we observe that since the function $g(\mathbf{A}\hat{\mathbf{A}}^\top \mathbf{z})$ is same, it suffices to work with the domain. We define the set $B = \{(\mathbf{A}\hat{\mathbf{A}}^\top)^{-1}\mathbf{z} : \mathbf{z} \in \mathcal{T}_{h,i}\}$. To obtain an upper bound for the near optimal cells in $\hat{\mathcal{T}}$ partitioning scheme, we consider every cell in $\hat{\mathcal{T}}$ partitioning scheme ($\hat{\mathcal{T}}_{h,i}$) that intersects with the B is potentially a near-optimal. To simplify our analysis, we enlarge

the domain B to make it a hypercube in $\mathbb{R}^m$. $\forall \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{T}_{h,i}$, consider,

$$
\begin{aligned}
&\left\|(\mathbf{A}\hat{\mathbf{A}}^\top)^{-1}\mathbf{z}_1 - (\mathbf{A}\hat{\mathbf{A}}^\top)^{-1}\mathbf{z}_1\right\|_\infty \\
&\leq \left\|(\mathbf{A}\hat{\mathbf{A}}^\top)^{-1}\right\|_\infty \|\mathbf{z}_1 - \mathbf{z}_2\|_\infty && \text{(From the Matrix Operator Norm definition)} \\
&\leq \sqrt{m}\left\|(\mathbf{A}\hat{\mathbf{A}}^\top)^{-1}\right\|_2 \|\mathbf{z}_1 - \mathbf{z}_2\|_\infty && \text{(From the Matrix Inequality 11)} \\
&\leq \frac{\sqrt{m}\left\|(\mathbf{A}\hat{\mathbf{A}}^\top)^{-1}\right\|_2 \|\mathbf{z}_1 - \mathbf{z}_2\|_\infty \sigma_{\min}(\mathbf{A}\hat{\mathbf{A}}^\top)}{\sigma_{\min}(\mathbf{A}\hat{\mathbf{A}}^\top)} && \text{(Since, } \sigma_{\min}(\mathbf{A}\hat{\mathbf{A}}) > 0) \\
&\leq \sqrt{m}\frac{\left\|(\mathbf{A}\hat{\mathbf{A}}^\top)^{-1}\mathbf{A}\hat{\mathbf{A}}^\top\right\|_2 \|\mathbf{z}_1 - \mathbf{z}_2\|_\infty}{\sigma_{\min}(\mathbf{A}\hat{\mathbf{A}}^\top)} && \text{(From the Matrix Inequality 12)} \\
&= \frac{\sqrt{m}\,\|\mathbf{z}_1 - \mathbf{z}_2\|_\infty}{\sigma_{\min}(\mathbf{A}\hat{\mathbf{A}}^\top)} && (\|\mathbf{I}\|_2 = 1) \\
&\overset{(1)}{\leq} \frac{\sqrt{m}}{\sigma_{\min}(\mathbf{A}\hat{\mathbf{A}}^\top)} 2\alpha 3^{-k} \\
&= 2\alpha 3^{-k}\frac{\sqrt{m}}{\sqrt{1 - \text{dist}^2(\mathbf{A}, \hat{\mathbf{A}})}} && \text{(From the Equation 8)}
\end{aligned}
$$

(1) is true from the following inequality,

$$
\begin{aligned}
\|\mathbf{z}_1 - \mathbf{z}_2\|_\infty &= \left\|\mathbf{s}\left(\frac{\mathbf{z}_1 - \mathbf{z}_{h,i} + \mathbf{z}_{h,i} - \mathbf{z}_2}{\mathbf{s}}\right)\right\|_\infty = \left\|\text{diag}\,(\mathbf{s})\left(\frac{\mathbf{z}_1 - \mathbf{z}_{h,i} + \mathbf{z}_{h,i} - \mathbf{z}_2}{\mathbf{s}}\right)\right\|_\infty \\
&\leq \|\text{diag}\,(\mathbf{s})\|_\infty \left\|\left(\frac{\mathbf{z}_1 - \mathbf{z}_{h,i} + \mathbf{z}_{h,i} - \mathbf{z}_2}{\mathbf{s}}\right)\right\|_\infty \leq 3^{-k}2\alpha
\end{aligned}
$$

Hence, $B \subseteq 2\alpha 3^{-k}\frac{\sqrt{m}}{\sqrt{1-\text{dist}^2(\mathbf{A},\hat{\mathbf{A}})}}\mathbb{H}_1^m$. To establish an upper bound for the near-optimal cells of $\hat{\mathcal{T}}$ partitioning scheme, we consider the region $2\alpha 3^{-k}\frac{\sqrt{m}}{\sqrt{1-\text{dist}^2(\mathbf{A},\hat{\mathbf{A}})}}\mathbb{H}_1^m$ as potentially near-optimal. We can then count the maximum number of cells in the $\hat{\mathcal{T}}$ partitioning scheme at height h that can intersect with this region.

To invoke the Lemma 19, we need the side-length of the cell $\hat{\mathcal{T}}_{h,i}$ cell or for simplicity, lower-bound to the side-length will also work. And the side-length is greater than $2 \cdot 3^{-(k+1)}\frac{\alpha\sqrt{m}}{\sqrt{1-\text{dist}^2(\mathbf{A},\hat{\mathbf{A}})}}$. Hence, the maximum number of cells of $\hat{\mathcal{T}}_{h,i}$ that can be tiled inside the hypercube with side length $= 2\alpha 3^{-k}\frac{\sqrt{m}}{\sqrt{1-\text{dist}^2(\mathbf{A},\hat{\mathbf{A}})}}$ are

$$
\left(1 + \left\lceil \frac{2\alpha 3^{-k}\frac{\sqrt{m}}{\sqrt{1-\text{dist}^2(\mathbf{A},\hat{\mathbf{A}})}}}{2 \cdot 3^{-(k+1)}\frac{\alpha\sqrt{m}}{\sqrt{1-\text{dist}^2(\mathbf{A},\hat{\mathbf{A}})}}} \right\rceil\right)^m = 4^m
$$

Hence, upper-bound to the near-optimal calls of $\hat{\mathcal{T}}$ partitioning scheme is

$$\forall h \geq 0, \quad \mathcal{N}_{\hat{\mathcal{T}}}(3\nu_{\mathcal{T}}\rho_{\mathcal{T}}^{h}) \leq 4^{m}\mathcal{N}_{\mathcal{T}}(3\nu_{\mathcal{T}}\rho_{\mathcal{T}}^{h}) \tag{45}$$

Using the above relation, we relate the near-optimality dimension. Suppose, $\eta_{\mathcal{T}}(\nu_{\mathcal{T}}, \rho_{\mathcal{T}}, C_{\mathcal{T}})$ is the near optimality dimension of $\mathcal{T}$ then,

$$\forall h \geq 0, \quad \mathcal{N}_{\mathcal{T}}(3\nu_{\mathcal{T}}\rho_{\mathcal{T}}^{h}) \leq C_{\mathcal{T}}\rho_{\mathcal{T}}^{-\eta_{\mathcal{T}}h} \tag{46}$$

From Lemma 24, we have the following SequOOL parameter relations:

$$\nu_{\hat{\mathcal{T}}} = \max\{\nu_{\mathcal{T}}, g^{*} - \inf_{\mathbf{z}\in\kappa_2\alpha\mathbb{H}_1^m} g(\mathbf{A}\hat{\mathbf{A}}^{\top}\mathbf{z})\}/\rho_{\mathcal{T}}^{\tilde{h}_2}, \rho_{\hat{\mathcal{T}}} = \rho_{\mathcal{T}}$$

$$\text{where } \tilde{h}_2 = m + m\left\lceil \log_3 \frac{2\sqrt{m}\kappa_2}{\kappa_2 - 1}\right\rceil \text{ and } \quad \kappa_2 = \frac{\sqrt{m}}{\sqrt{1 - \mathrm{dist}^2\left(\mathbf{A}, \hat{\mathbf{A}}\right)}}.$$

For brevity, denote $\nu'_{\mathcal{T}} = \max\{\nu_{\mathcal{T}}, g^{*} - \inf_{\mathbf{z}\in\kappa_2\alpha\mathbb{H}_1^m} g(\mathbf{A}\hat{\mathbf{A}}^{\top}\mathbf{z})\}$. Now, consider $3\nu_{\hat{\mathcal{T}}}\rho_{\hat{\mathcal{T}}}^{h}$:

$$\begin{aligned}
&= 3\nu'_{\mathcal{T}}\rho_{\mathcal{T}}^{h-\tilde{h}_2} \\
&= 3\nu_{\mathcal{T}}\rho_{\mathcal{T}}^{\log_{\rho_{\mathcal{T}}}(\nu'_{\mathcal{T}}\rho_{\mathcal{T}}^{-\tilde{h}_2}) - \log_{\rho_{\mathcal{T}}}(\nu_{\mathcal{T}})}\rho_{\mathcal{T}}^{h} \\
&\leq 3\nu_{\mathcal{T}}\rho_{\mathcal{T}}^{\left\lfloor \log_{\rho_{\mathcal{T}}}(\nu'_{\mathcal{T}}\rho_{\mathcal{T}}^{-\tilde{h}_2}) - \log_{\rho_{\mathcal{T}}}(\nu_{\mathcal{T}})\right\rfloor}\rho_{\mathcal{T}}^{h} \\
&= 3\nu_{\mathcal{T}}\rho_{\mathcal{T}}^{-\tilde{h}_4}\rho_{\mathcal{T}}^{h} && \text{(Denote } \tilde{h}_4 = -\left\lfloor \log_{\rho_{\mathcal{T}}}(\nu'_{\mathcal{T}}\rho_{\mathcal{T}}^{-\tilde{h}_2}) - \log_{\rho_{\mathcal{T}}}(\nu_{\mathcal{T}})\right\rfloor)
\end{aligned}$$

Next, we examine $\mathcal{N}_{\hat{\mathcal{T}}}(3\nu_{\hat{\mathcal{T}}}\rho_{\hat{\mathcal{T}}}^{h})$:

$$\begin{aligned}
&\leq \mathcal{N}_{\hat{\mathcal{T}}}(3\nu_{\mathcal{T}}\rho_{\mathcal{T}}^{h-\tilde{h}_4}) \\
&\leq 4^{m}\mathcal{N}_{\mathcal{T}}(3\nu_{\mathcal{T}}\rho_{\mathcal{T}}^{h-\tilde{h}_4}) && \text{(Using Inequality 46)} \\
&\leq 4^{m}C_{\mathcal{T}}\rho_{\mathcal{T}}^{-\eta_{\mathcal{T}}(h-\tilde{h}_4)} && \text{(Using Inequality 45 and } \forall h \geq \tilde{h}_4)
\end{aligned}$$

Therefore, we have:

$$\mathcal{N}_{\hat{\mathcal{T}}}(3\nu_{\hat{\mathcal{T}}}\rho_{\hat{\mathcal{T}}}^{h}) \leq 4^{m}C_{\mathcal{T}}\rho_{\mathcal{T}}^{\eta_{\mathcal{T}}\tilde{h}_4}\rho_{\mathcal{T}}^{-\eta_{\mathcal{T}}h} \quad \forall h \geq \tilde{h}_4 \tag{47}$$

For heights $h \in [0 : \tilde{h}_4 - 1]$, we can use the value of the right-hand side at depth $h = \tilde{h}_4$, which is $4^{m}C_{\mathcal{T}}$. Hence, we have:

$$\begin{aligned}
\mathcal{N}_{\hat{\mathcal{T}}}(3\nu_{\hat{\mathcal{T}}}\rho_{\hat{\mathcal{T}}}^{h}) &\leq 4^{m}C_{\mathcal{T}}\rho_{\mathcal{T}}^{\eta_{\mathcal{T}}\tilde{h}_4}\rho_{\mathcal{T}}^{-\eta_{\mathcal{T}}h} \quad \forall h \geq 0 \\
&= 4^{m}C_{\mathcal{T}}\rho_{\mathcal{T}}^{\eta_{\mathcal{T}}\tilde{h}_4}\rho_{\hat{\mathcal{T}}}^{-\eta_{\mathcal{T}}h} \quad \forall h \geq 0
\end{aligned}$$

Therefore, we conclude that $\eta_{\hat{\mathcal{T}}} \leq \eta_{\mathcal{T}}$ and $C_{\hat{\mathcal{T}}} = C_{\mathcal{T}}4^{m}\rho_{\mathcal{T}}^{\eta_{\mathcal{T}}\tilde{h}_4}$. ■

F.18. Proof of Proposition 7

Proof

Let $h_{\max} = \left\lfloor \frac{n^2}{n\overline{\log n} + \frac{Tn}{3c}} \right\rfloor$ as given in the lemma statement. SequOOL opens $\left\lfloor \frac{h_{\max}}{h} \right\rfloor$ cells at depth h for all $h \in [1, h_{\max}]$. Additionally, we utilize T samples for every c heights to learn $\hat{f}$. The total number of samples used to learn $\hat{f}$ up to height $h_{\max}$ is thus $Th_{\max}/c$. Each cell opening in SequOOL requires 3 samples. Therefore, the total number of openings performed by Algorithm 3 is given by $\sum_{i=1}^{h_{\max}} \left\lfloor \frac{h_{\max}}{i} \right\rfloor + \frac{Th_{\max}}{3c}$. According to the proposition, we need to show that this quantity is $\leq n$. Consider, total number of openings:

$$
\begin{aligned}
&= \sum_{i=1}^{n} \left\lfloor \frac{h_{\max}}{i} \right\rfloor + \frac{h_{\max} T}{3c} \\
&\leq \left\lfloor \frac{n^2}{n\overline{\log n} + \frac{Tn}{3c}} \right\rfloor \left(\sum_{i=1}^{n} \frac{1}{i} + \frac{T}{3c}\right) && \text{(Definition of } h_{\max}\text{)} \\
&\leq \frac{n^2}{n\overline{\log n} + \frac{Tn}{3c}} \left(\overline{\log n} + \frac{T}{3c}\right) && \text{(Recall the definition: } \overline{\log n} \triangleq \textstyle\sum_{t=1}^{n} \frac{1}{t}\text{)} \\
&= n
\end{aligned}
$$

Hence, the number of openings made in Algorithm 3 does not exceed n.

F.19. Proof of Theorem 27

We start with the Theorem 5 of [1]. We restate the theorem, adapting it to our notation and incorporating the dependency of the parameters on the partitioning scheme $\mathcal{P}$:

Theorem 34 ([1], Theorem 5) *Let W be the standard Lambert W function. Suppose f along the partitioning scheme $\mathcal{P}$ satisfies Assumption 12 with associated $(\nu_{\mathcal{P}}, \rho_{\mathcal{P}})$, $C_{\mathcal{P}} > 1$, and near-optimality dimension $\eta_{\mathcal{P}} = \eta_{\mathcal{P}}(\nu_{\mathcal{P}}, C_{\mathcal{P}}, \rho_{\mathcal{P}})$ parameters. Then, after n rounds, the simple regret of SequOOL is bounded as follows: For $\eta_{\mathcal{P}} > 0$, we use Corollary 6 of [1]. Let $\tilde{n} = \left\lfloor n/\overline{\log n} \right\rfloor \eta_{\mathcal{P}} \log(1/\rho_{\mathcal{P}})/(C_{\mathcal{P}})$.*

- *If $\eta_{\mathcal{P}} = 0, r_n \leq \nu_{\mathcal{P}} \rho_{\mathcal{P}}^{\frac{1}{C_{\mathcal{P}}} \left\lfloor \frac{n}{\overline{\log n}} \right\rfloor}$*
- *If $\eta_{\mathcal{P}} > 0, r_n \leq \nu_{\mathcal{P}} \left(\frac{\tilde{n}}{\overline{\log \tilde{n}}}\right)^{-\frac{1}{\eta_{\mathcal{P}}}}$*

To invoke this Theorem for our proof, first we apply the theorem for the partitioning scheme $\hat{\mathcal{A}}$. 25 shows that $\hat{\mathcal{A}}$ is a valid partioning scheme, i.e., it satisfies 12, hence we can invoke Theorem 34.

Thus, for our partitioning scheme $\hat{\mathcal{A}}$, denoting $\tilde{n} = \left\lfloor n/\overline{\log n} \right\rfloor \eta_{\hat{\mathcal{A}}} \log(1/\rho_{\hat{\mathcal{A}}})/(C_{\hat{\mathcal{A}}})$, the regret is bounded by

- If $\eta_{\hat{\mathcal{A}}} = 0, r_n \leq \nu_{\hat{\mathcal{A}}} \rho_{\hat{\mathcal{A}}}^{\frac{1}{C_{\hat{\mathcal{A}}}} \left\lfloor \frac{n}{\overline{\log n}} \right\rfloor}$
- If $\eta_{\hat{\mathcal{A}}} > 0, r_n \leq \nu_{\hat{\mathcal{A}}} \left(\frac{\tilde{n}}{\overline{\log \tilde{n}}}\right)^{-\frac{1}{\eta_{\hat{\mathcal{A}}}}}$

Corollary 26 relates SequOOL parameters and gives,

$$
\begin{aligned}
&\rho_{\hat{\mathcal{A}}} = \rho^{\beta}, \nu_{\hat{\mathcal{A}}} = \max\{\max\{\nu, l_f\}\rho^{(1-\beta)(m-1)-\tilde{h}_1}, l_g\}\rho_{\mathcal{A}}^{-\tilde{h}_2}, \\
&\eta_{\hat{\mathcal{A}}}(\nu_{\hat{\mathcal{A}}}, \rho_{\hat{\mathcal{A}}}, C_{\hat{\mathcal{A}}}) \leq \frac{\eta_{\mathcal{P}}(\nu, \rho, C)}{\beta}, C_{\hat{\mathcal{A}}} = 3^d d^{d-m} (12\sqrt{m})^m C \rho^{-\eta_{\mathcal{P}} \tilde{h}_3} 4^m \rho^{\eta_{\mathcal{P}} \tilde{h}_4}.
\end{aligned}
$$

Now, we substitute these relations in our regret bound to get the upper bound in terms of default partitioning scheme $\mathcal{P}$ parameters.

$$r_n \leq \begin{cases} \max\{\max\{\nu, l_f\}\rho^{(1-\beta)(m-1)-\tilde{h}_1}, l_g\}\rho^{-\beta\tilde{h}_2}\rho^{\frac{\beta}{C_1}\lfloor\frac{n}{\overline{\log n}}\rfloor} & \text{if } \eta_{\mathcal{P}} = 0, \\ \max\{\max\{\nu, l_f\}\rho^{(1-\beta)(m-1)-\tilde{h}_1}, l_g\}\rho^{-\beta\tilde{h}_2}\left(\frac{\tilde{n}}{\overline{\log \tilde{n}}}\right)^{-\frac{\beta}{\eta_{\mathcal{P}}}} & \text{if } \eta_{\mathcal{P}} > 0, \end{cases}$$

Where $C_1 = 3^d d^{d-m}(12\sqrt{m})^m C\rho^{-\eta_{\mathcal{P}}\tilde{h}_3}4^m\rho^{\eta_{\mathcal{P}}\tilde{h}_4}$ and $\tilde{n} = \lfloor n/\overline{\log n}\rfloor\, \eta_{\mathcal{P}}\log(1/\rho)/C_1$ with $\tilde{h}_2 = m + m\left\lceil\log_3 \frac{2\sqrt{m}\kappa_2}{\kappa_2-1}\right\rceil$ and $\tilde{h}_1 = d\lceil\log_3\sqrt{m}\alpha\rceil$ with $\kappa_2 = \frac{\sqrt{m}}{\sqrt{1-\text{dist}^2(\mathbf{A},\hat{\mathbf{A}})}}$

$$\tilde{h}_3 = -\left\lfloor\log_\rho(\max\{\nu, l_f\}\rho^{(1-\beta)(m-1)-\tilde{h}_1}) - \log_\rho(\nu)\right\rfloor$$
$$\tilde{h}_4 = -\left\lfloor\log_{\rho_{\mathcal{A}}}(\max\{\nu_{\mathcal{A}}, l_g\}\rho_{\mathcal{A}}^{-\tilde{h}_2}) - \log_{\rho_{\mathcal{A}}}(\nu_{\mathcal{A}})\right\rfloor$$

■

Appendix G. Additional Experiment Details & Results

G.1. Test Functions Experiments

We implemented SequOOL, SOO, and RESOO ourselves due to the absence of publicly available open-source code for these algorithms. For DiRect and Dual Annealing, we utilized the implementations provided in the SciPy library's optimize module. The CMA-ES algorithm was sourced from its dedicated project repository[1]. REMBO and HesBO implementations were derived from the original HesBO repository[2].

G.2. Multi-Index Functions Results

We present additional experimental results to further demonstrate the effectiveness of our approach. Figure 4 showcases the performance of various algorithms on low-dimensional multi-index functions with $d = 5$ and $m = 2$. Our algorithm consistently achieves lower regret across different test functions, including Sphere, Branin, Ellipsoid, and Rastrigin, often reaching zero regret with fewer samples compared to competing methods.

G.3. LLM Quantization

The AWQ [13] method for quantizing large language models formulates optimization problem as: $\alpha^* = \arg\min_{\alpha\in[0,1]}\mathcal{L}(\mathbf{s}_X^\alpha)$, where $\mathcal{L}(\mathbf{s}) = \left\|Q(\mathbf{W}\cdot\mathbf{s})(\mathbf{s}^{-1}\cdot\mathbf{X}) - \mathbf{W}\mathbf{X}\right\|_2$. Where $\mathbf{X}$ is the input features to the block which is cached from a calibration dataset. It uses the parameterization $\mathbf{s} = \mathbf{s}_\mathbf{X}^\alpha$, where $\mathbf{s}_\mathbf{X}$ is the activation

Table 2: LLM Quantization Experiment Results

Algorithm	Compute Time	WikiText-2 PPL	Calibration -Set PPL
Grid Search	≈ 9 hours	16.92	14.62
SequOOL	≈ 10 hours	16.83	**14.28**
Algorithm 1	≈ 10 hours	16.96	14.42
Algorithm 2	≈ 12 hours	**16.68**	**14.29**

1. https://github.com/CyberAgentAILab/cmaes
2. https://github.com/aminnayebi/HesBO

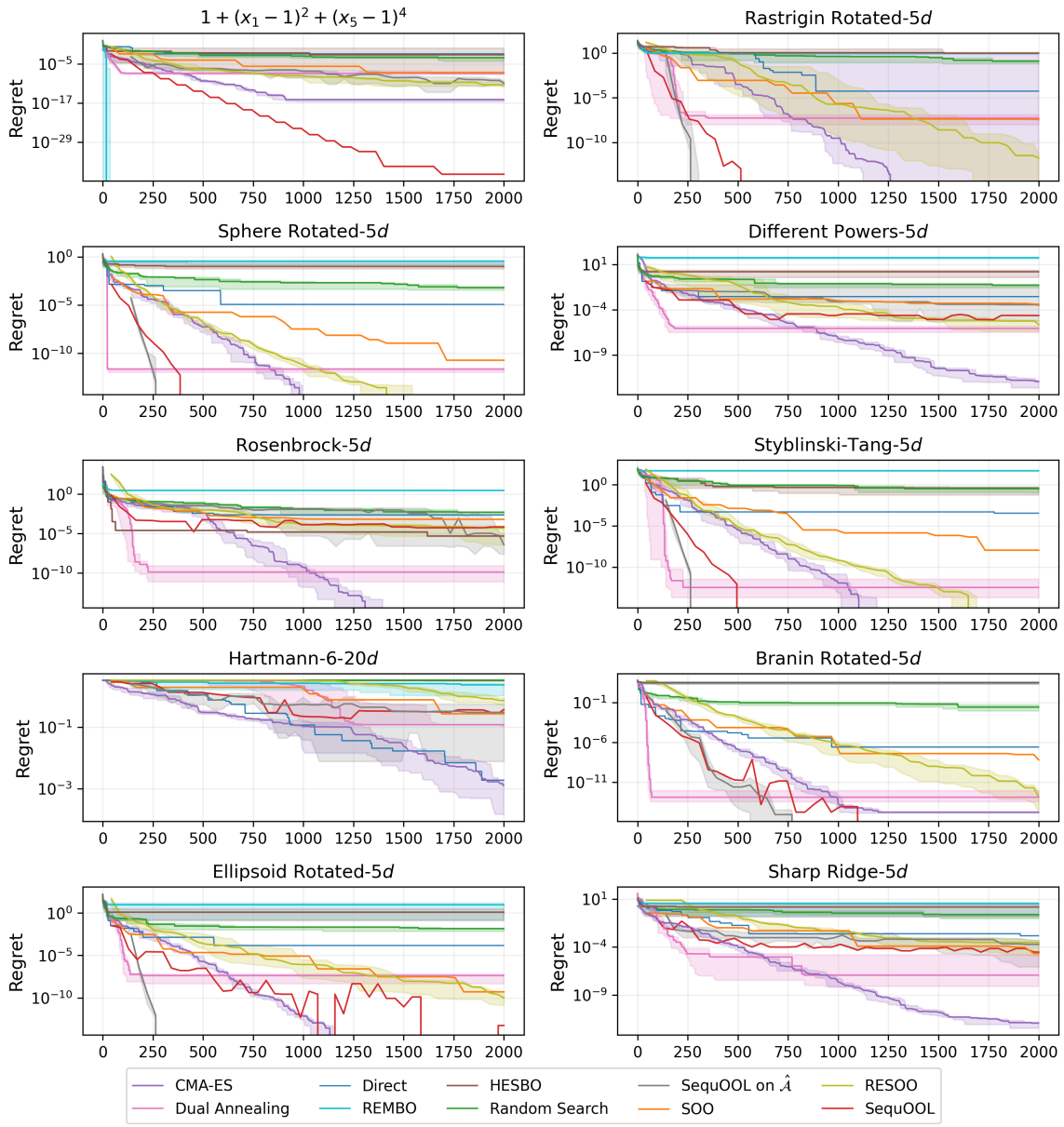

Figure 4: Regret Plots: Legend for the plots are arranged in the order of their performance. Algorithm 1 (SequOOL on $\hat{\mathcal{A}}$) uses 100 additional samples to learn the subspace through the Fornasier et al. [5] approach. Our Algorithm, SOO, RESOO, SequOOL are budget algorithms, so we run these algorithms using 100 equally spaced budget values between 1 and 2000 and plot the regret at the end of each run. For the randomized algorithms, we took 10 trials and plotted the median curve (thick line) and 0 and 95 percentile curves.

scale computed from $\mathbf{X}$ and $\alpha \in [0, 1]$ and Q as the quantization function and $\mathbf{W}$ as the original weights. And in their work α^* is found through 1D grid search.

To optimize this single parameter α , the method utilizes a grid search approach for each of the three primary components in every layer of the large language model: the attention matrices ($\mathbf{W}^Q$, $\mathbf{W}^K$, $\mathbf{W}^V$, and $\mathbf{W}^O$: All the four matrices have one parameter α), the first fully connected layer ($\mathbf{W}^{\texttt{fc1}}$), and the second fully connected layer ($\mathbf{W}^{\texttt{fc2}}$). Consequently, this leads to three optimization parameters per layer, resulting in a total of 3M parameters to optimize across M layers. Each of these parameters is derived from separate optimization problems, all of which are solved through the grid search method in the interval $[0, 1]$ to find the optimal value of α^*.

We propose a new approach which involves solving this LLM Quantization as high-dimensional black-box optimization problem. In our approach, we jointly optimize all layers to minimize perplexity: So, our approach has one optimization problem in $3M$ dimensional space, compared to AWQ which has $3M$ one-dimensional optimization problems. Let $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_{3M}]^\top$ represent the scales for all M layers, with each layer having three parameters. We define our proposed optimization problem as: $\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \mathcal{P}(\boldsymbol{\alpha})$ subject to $\alpha_i \in [0, 1] \quad \forall i \in \{1, ..., 3M\}$. Where, $\mathcal{P}(\boldsymbol{\alpha})$ is the perplexity on the calibration set after quantization using the scaling factors derived from $\boldsymbol{\alpha}$.

We evaluated our approach on the OPT-1.3B [31] model and compared it with AWQ in Table 2. Our proposed objective function using SequOOL over 72 dimensions outperformed AWQ, achieving lower perplexity on both WikiText-2 [15] and the calibration set (Pile dataset [6]). More details are in Appendix G.4.

G.4. Training Details of LLM Quantization Experiment

We implemented our Large Language Model (LLM) code on hardware equipped with one Quadro RTX 5000 GPU having 16GB VRAM. For comparison, we ran AWQ baselines using the original authors' code, which also served as a foundation for developing our proposed method.

To optimize the neural network used in Algorithm 3 for our LLM Quantization objective function, we employed the Ray package for hyper-parameter tuning [3]. We used Adam optimizer and our search space included hidden layer sizes (500, 1000, 2000, 3000), learning rates (log-uniform from 1×10^{-4} to 1×10^{-1}), weight decay (log-uniform from 1×10^{-2} to 1×10^{-1}), and learning rate Step Decay with gamma values (uniform from 0.9 to 0.99), and step sizes (500, 1000, 2000). We utilized early stopping to prevent overfitting.

The neural network was retrained on SequOOL-collected samples every 5 heights, with the look-ahead strategy applied up to a height of 60 and performed round-robin direction selection after this height.

3. https://github.com/ray-project/ray