ABSTRACT

Title of Dissertation: CONVERGENCE RATE THEORY FOR GLOBAL OPTIMIZATION

Jialin Li
Doctor of Philosophy, 2021

Dissertation Directed by: Professor Ilya Ryzhov
Department of Decision, Operations, and Information Technologies

Global optimization is used to control complex systems whose response is an unknown function on a continuous domain. Response values can only be observed empirically by simulations, and cannot be accurately represented using closed-form mathematical expressions. Prediction of true optimizer in this context is usually accomplished by constructing a surrogate model that can be thought of as an interpolation of a discrete set of observed design points.

This thesis includes study of convergence rates of epsilon-greedy global optimization under radial basis function interpolation. We derive both convergence rates and concentration inequalities for a general and widely used class of interpolation models known as radial basis functions, used in conjunction with a randomized algorithm that searches for solutions either within a small neighborhood of the current-best, or randomly over the entire domain. An interesting insight of this work is that the convergence rate is improved when the size of the local search region shrinks to zero over time in a certain way. My work precisely characterizes the rate of this shrinkage.

Gaussian process regression is another tool that is widely used to construct surrogate models. A theoretical framework is developed for proving new moderate deviations inequalities on different types of error probabilities that arise in GP regression. Two specific examples of broad interest are the probability of falsely ordering pairs of points (incorrectly estimating one point as being better than another) and the tail probability of the estimation error of the minimum value. Our inequalities connect these probabilities to the mesh norm, which measures how well the design points fill the space. Convergence rates are further instantiated in settings of using a Gaussian kernel, and either deterministic or random design sequences.

Convergence can be more rapid when we are not totally blind to the objective function. As an example, we present a work on simultaneous asymmetric orthogonal tensor decomposition. Tensor decomposition can be essentially viewed as a global optimization problem. However with the knowledge of the algebraic information from the observed tensor, the method only requires $O(\log(\log \frac{1}{\epsilon}))$ iterations to reach a precision of ϵ.

CONVERGENCE RATE THEORY FOR GLOBAL OPTIMIZATION

by

Jialin Li

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Ilya Ryzhov, Chair/Advisor
Professor Michael Fu
Professor Eric Slud
Professor Leonid Koralov
Professor André Tits

Acknowledgments

Me getting a Ph. D. degree is about a process of me getting a deeper understanding of myself. For many times I read people who had a clear picture of their long-term goal at youth, and I realized that I don't have the fortune to be one of them. But looking back, I have the fortune to meet professor Ryzhov.

I owe my greatest thanks to my advisor professor Ilya Ryzhov, who does what a good advisor does. He guided my tour to the beautiful quantitative world, shared his research taste and educative insights, fed me tricky but neat research problems, encouraged me to continue a longer academic journey, and finally helped me to make a first step on that career path. For this thesis to come into reality, I appreciate that he left me enough space to explore the topics I like, and offered inspiring ideas at key moments.

I would also like to thank professor Michael Fu, professor Eric Slud, professor Leonid Koralov and professor André Tits for their acceptance to serve as my committee of dissertation defense, their valuable time and effort in examining the quality of my work.

It is the treasure of my Ph. D. years to make acquaintance with my dearest friends. They are important to me not only for the company and laugh, but also because from time to time I can always learn something from them. Their being gives me motivation to jump higher.

I thank the people who refered me to a job or wrote me a recommendation letter

during my job-searching phase. I thank my parents who gave birth to me so that I can experience joy and depression. I thank all the kind and interesting people I met over the past 5 years. I wish everybody stay well and healthy, and be happy about their job and lives.

In the end, please allow me to thank the persistance of myself in chasing for the things I value, and the existance of those philosophical questions I don't yet have an answer for.

Table of Contents

Chapter 1: Introduction

Optimization is one of the most important quantitative problems in the world and it is so close to affect our daily lives. The history and development of optimization is comprised of a board range of subjects. From statistics and operations research to computer science and engineering, the logic is alike - to minimize or maximize the response of a complex system whenever we can model the system as a real function of the configurations we could make. The only difference over these subjects is that they characterize the unknown under various background information and identify problem-specific input domains. There are a number of branches under this general topic, as one can find papers for single-objective or multi-objective optimizations, for categorical or continuous domains, and for local or global optimization (unique or multiple extrema).

We will particularly focus on global optimization for single-objective settings in this thesis. Global optimization is used to control multiple-extrema systems whose performance can only be observed empirically, by running physics-based or stochastic simulations, and cannot be accurately represented using closed-form mathematical expressions. For example, battery electric vehicle engineers use simulation-based optimization to identify the most effective design variables and compare different configurations of various car segments. Similar “knob tuning” is also used to select parameters for drilling

new oil wells, or even to improve the predictive performance of high-dimensional neural networks. In all of these cases, one can evaluate the performance of a particular configuration of parameters through a black-box simulator (or through a field experiment), but each such evaluation is computationally expensive. Thus, exhaustive search is impossible, and one has to use the results of a small number of experiments to accurately predict the outcomes of others.

Prediction is usually accomplished by constructing a surrogate model (can be thought of as an interpolation of the observed solutions). Using this model, one can make an educated guess as to the best configuration, and potentially run additional experiments based on this information. The final result thus depends on two factors - the specific technique used to construct the interpolation, and the logic used to design new experiments. Researchers have developed a rich set of tools for both of these aspects, with most of the literature focusing on their computational performance.

The theory behind these algorithms, however, is less developed, especially when it comes to convergence rates, which rigorously quantify how much "effort" is needed to solve a given problem with a certain degree of precision. The second chapter of this thesis is one of a very small number of studies to investigate this issue. Both convergence rates and concentration inequalities are derived for a general and widely used class of interpolation models known as "radial basis functions," used in conjunction with a randomized algorithm that searches for solutions either within a small neighborhood of the current-best, or randomly over the entire domain. This algorithm is simple, but theoretically tractable, and it captures the basic tradeoff between local and global search that is fundamental to virtually any global optimization algorithm.

An interesting insight of this work is that the convergence rate is improved when the size of the local search region shrinks to zero over time in a certain way. My paper precisely characterizes the rate of this shrinkage - a slower rate will waste effort that could have been used to explore the rest of the domain, while a faster rate runs the risk of stalling at a suboptimal solution. On the other hand, the weight assigned to global search should always be positively lower bounded, i.e., it is never safe to stop exploring the domain entirely.

Gaussian process (GP) regression can be another tool that is widely used to construct surrogate models. This is a classical model for geostatistics, Bayesian optimization, and parameter tuning in machine learning. It also provides another perspective with stochastic components to the radial basis function interpolation method, although there might not be any change in our implementation empirically. So this part can be viewed as a complementary work to the foremost convergence rate analysis under radial basis function interpolation.

Chapter 3 contains a theoretical framework that is developed for proving new moderate deviations inequalities on different types of error probabilities that arise in GP regression. Two specific examples of broad interest are the probability of falsely ordering pairs of points (incorrectly estimating one point as being better than another) and the tail probability of the estimation error of the minimum value. Our inequalities connect these probabilities to the mesh norm, which measures how well the design points fill the space. Convergence rates are further instantiated in settings of using a Gaussian kernel, and either deterministic or random design sequences.

The description of the term precision can be an interesting topic. Connecting chap-

ter 2 and chapter 3, we see that when researcher seeks for less, or weaker precision, the convergence rate can be improved. The result of chapter 2 shows that if one cares how well the estimated minimizer in the domain approaches to the true point, the convergence rate can be polynomial. This can be compared with the work in chapter 3 where a Gaussian process prior is adopted to account for the randomness of responses. As we may use the same formula to construct surrogates, the result from chapter 2 still works for chapter 3. However the Gaussian model provides space for studying the convergence of the probability of the extreme value estimation having an error larger than a preset threshold. This convergence turns out to be on an exponential order.

More generally, if any additional information are there to help so that the objective is not any more a pure black box, we will have another dimension to improve the convergence rate of global optimization. In this thesis, the last major work on tensor decomposition can serve as a critical instance to support that argument. Decomposing a tensor, although not necessarily based on any general optimization logic, is indeed equivalent to a global optimization problem - the components are searched and adjusted to best recover the observed tensor. Reader of this thesis will see that, by taking advantage of the algebraic structure of the tensor data, we will be able to achieve a rapid convergence even faster than exponential decays.

The work on simultaneous asymmetric tensor decomposition via alternating subspace iteration is presented in chapter 4. Tensor decomposition has a long history connected with many scientific disciplines such as psychometrics and neuroscience. Benefited from rapid development on hardware in the last decade, the significant interest in this multi-way data is reflected in emerging engineering applications in biomedical area, sig-

nal processing, data mining and computer vision. Specifically, for latent variable models (including Gaussian mixture models, hidden Markov models, latent Dirichlet allocation, etc.), tensor decomposition can be used to develop estimators by method of moments. The method of moments is simple as first compute the tensor of empirical moments as sample means and correlations, then solve for the model parameters that give rise to (nearly) the observed quantities. This constructive method leads to consistent estimators which can be efficiently computed by orthogonally decomposing a tensor of observed moments. This efficiency becomes especially valuable in a high-dimension problem since the number of cross-feature moments can be large.

In particular, we pay attention on a specific kind of tensor decomposition called tensor CP decomposition. Existing popular approaches either recover components one by one, not necessarily in the order of larger components first, or requires matrix decomposition which requires a linear convergence rate. Recently developed simultaneous power method, although achieves a quadratic convergence rate, obtains only a high probability recovery of top r components even when the observed tensor is noiseless. For the purpose of improving computational efficiency, a new algorithm is developed for decomposition that is able to handle asymmetric tensors.

We propose a Slicing Initialized Alternating Subspace Iteration method and a Slice-Based Initialization procedure that together guarantee the almost sure recovery of top r components (ϵ-close) under noiseless cases. When tensor is noisy, our algorithm is provably robust to noise and has high probability to achieve the goal. The alternating subspace iteration method runs $O(\log(\log \frac{1}{\epsilon}))$ steps of tensor subspace iterations while the initialization takes only constant steps of matrix subspace iterations, which is a sign

of efficiency.

Typically in this literature, the eigenvectors characterizing the decomposition are found sequentially; the state of the art is able to recover them simultaneously, but only for the symmetric case. My work is the first to guarantee simultaneous recovery in the asymmetric setting, with rigorous convergence rates that hold even if the tensor rank is misspecified. Furthermore, under the noiseless case my approach is guaranteed to converge almost surely, covering both symmetric and asymmetric tensors, which is particularly notable because many prior methods in this area can only be proved to succeed with high probability (not necessarily 1). When tensor is noisy, our algorithm is provably robust to noise and has high probability to simultaneously recover top components.

Chapter 2: Convergence Rates of Epsilon-Greedy Global Optimization Under Radial Basis Function Interpolation

2.1 Introduction

Consider the optimization problem $\min_{x \in \mathcal{X}} f(x)$, where no explicit-form expression for f is available. We can observe the function values $f(x_n)$ at individual *design points* $\{x_n\}_{n \geq 1}$ of our choosing (we assume that the observations are noiseless). f has certain smoothness but we have no information about the derivative of f at these points. This problem class is also known as "global optimization" and "derivative-free optimization" [Conn et al., 2009a], and is often applied to tune parameters in engineering simulators [Giuliani and Camponogara, 2015] or machine learning models [Eitrich and Lang, 2006].

When the domain $\mathcal{X} \subseteq \mathbb{R}^d$ is compact and connected, at stage N we construct a function $\hat{f}_N$ (often called a "metamodel" or "surrogate model") that interpolates the observed function values $f(x_1), ..., f(x_N)$ in some way. This allows us to predict values at points we have not yet observed, and to approximate the optimal solution $x^* = \arg\min_{x \in \mathcal{X}} f(x)$ (assumed to be unique) by calculating $\hat{x}_N^* = \arg\min_{x \in \mathcal{X}} \hat{f}_N(x)$ (randomly selected if not unique). We can also use the interpolation to guide the selection

of new design points, i.e., x_N can be allowed to depend on $\hat{f}_{N-1}$. Thus, the quality of our estimate of the optimal solution (i.e., the difference $\|\hat{x}_N^* - x^*\|$) is determined by two factors: 1) the particular interpolation method used to construct $\hat{f}_N$, and 2) the policy used to determine $\{x_n\}$ based on previous observations.

In this chapter, we derive new results on the *convergence rate* of $\|\hat{x}_N^* - x^*\|$ under specific choices for the policy and interpolation. We assume that the metamodel $\hat{f}_N$ is constructed using the method of radial basis functions (RBFs), which is widely used in global optimization, and is closely related to Gaussian process regression (itself a very popular technique). As for the choice of design points, we focus on a relatively simple sequential policy known as ε-greedy: at each time stage n, we either sample uniformly from a small neighborhood of $\hat{x}_{n-1}^*$ (with probability $\varepsilon > 0$), or we sample uniformly from the entire domain $\mathcal{X}$ (with probability $1 - \varepsilon > 0$); note that this policy is randomized. We will explain the reasons for this choice of policy further down, but first we will state the two main results of this chapter: the pathwise convergence rate

$$\|\hat{x}_N^* - x^*\| = O\left(\left(\frac{\log N}{N}\right)^{\frac{k}{2d}} \left(\frac{\log(bN)}{bN}\right)^{\frac{k^2}{4d}}\right) \qquad \text{a.s.},$$

where b is determined from ε (will later explain how) and k is a parameter of the interpolation model that can be computed by user, and the concentration inequality

$$P\left(\|\hat{x}_N^* - x^*\| > c\left(\frac{\log N}{N}\right)^{\frac{k}{2d}} \left(\frac{\log(bN)}{bN}\right)^{\frac{k^2}{4d}}\right) \leq \frac{c'}{N},$$

where c, c' are problem-specific constants. The pathwise rate is asymptotic, but at a fixed

time N, the set of all sample paths that have not yet entered the asymptotic regime has measure $O\left(\frac{1}{N}\right)$.

Although the global optimization literature has a long history, and many sophisticated sampling procedures have been developed, results of the above type remain fairly rare: most studies focus on empirical performance and/or on weaker theoretical guarantees such as convergence to a first-order critical point. Among papers that do study convergence rates, many require additional structure on f, such as convexity [Bauschke et al., 2015, Duchi et al., 2015] or strong convexity [Berahas et al., 2019]. Among the very few papers that do *not* require such assumptions, we highlight Bull [2011], which studies the a.s. convergence rate of $\mathbb{E}\left|f\left(\hat{x}_N^*\right) - f\left(x^*\right)\right|$ under Gaussian process interpolation and the expected improvement sampling procedure; the rate obtained is similar to ours. We also mention the recent work by Calvin et al. [2018], which obtains a very strong rate of $O\left(e^{-c\sqrt{N}}\right)$ on the optimality gap $\min_{1\leq n\leq N} f\left(x_n\right) - f\left(x^*\right)$, but requires a computationally expensive multilinear interpolation model as well as the numerical evaluation of complicated integrals; the constant c also vanishes very quickly in the dimension d. Lastly, Tikhomirov [2006] derives a bound on the time required to reach a certain accuracy using randomized direct search (without any interpolation model).

In light of this, there is value in focusing on the ε-greedy policy, which has had a long history in reinforcement learning [Sutton and Barto, 2018] and is still actively used in applications such as recommender systems [Kamishima and Akaho, 2011] and crowdsourcing [Raykar and Agrawal, 2014]. This simple policy captures the key tradeoff between local and global search, governed by the parameter ε. It enables a tractable analysis of convergence rates under RBF interpolation, and potentially would be scalable

to high-dimensional problems where more sophisticated methods run into computational bottlenecks. Furthermore, the rates that we derive also hold for generalizations of ε-greedy where global search is conducted by sampling from an arbitrary density on $\mathcal{X}$ (this changes the multiplicative constant, but not the order of the rate), so in that sense our choice of policy is not restrictive. However, our analysis cannot *improve* the rates by using non-uniform sampling, because our proof technique relies on a connection between the estimation error $\|\hat{f}_N - f\|$ under RBFs and the so-called "mesh norm," which measures how evenly the design points are spread out over $\mathcal{X}$.

We do, however, obtain an insight into the optimal size of the local search region. Namely, we find that the size of the neighborhood around $\hat{x}_N^*$ should *shrink* over time, at a rate proportional to $\left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}}$. Essentially, if the local search region shrinks too slowly, we will be wasting design points that should have been used to explore the domain; however, if the local search region shrinks too quickly, there is a risk that it will no longer cover x^* even when N is very large. Shrinking the local search region at the rate indicated above improves the convergence rate to the one aforementioned by a factor of $\left(\frac{\log(bN)}{bN}\right)^{\frac{k^2}{4d}}$ that otherwise would not be there.

2.2 Literature review

There is a large class of global optimization methods that either do not require a metamodel at all, or can be applied very generally (with virtually any metamodel). These include heuristics such as evolutionary algorithms [Back, 1996], simulated annealing [Corana et al., 1987] and particle swarm [Hu et al., 2004] algorithms. Such approaches

have shown promise in global optimization (see, e.g., Schutte and Groenwold, 2005 or Yang, 2010). To give some examples of the available theory, Van den Bergh and Engelbrecht [2006] proved convergence of particle swarm to stationary points, while Vaz and Vicente [2007] proved the existence of a subsequence of design points that converges to a first-order critical point. Orosz and Jacobson [2002] studied the expected number of samples required by simulated annealing to identify a suboptimal solution within some fixed tolerance level; the resulting bounds, however, are difficult to compute and have to be evaluated numerically.

Direct search methods [Torczon, 1997] also do not require an interpolation model, but rather move toward x^* by a sequence of local directional searches. This methodology can handle extensions such as constrained problems [Lewis and Torczon, 1999, 2000]; see Kolda et al. [2003] for a review of various extensions and improvements. A major advance in this literature was the development of mesh-adaptive direct search [Audet and Dennis, 2006], which allows substantially more flexibility in the choice of direction. Again, many extensions are possible, for example to nonsmooth optimization [Audet et al., 2008] or multiobjective optimization [Audet et al., 2010]. The theory generally focuses on convergence to first-order critical points, with Abramson and Audet [2006] proving convergence to second-order stationary points.

Trust-region methods conduct local search on a suitably defined region using a metamodel, for example linear [Powell, 1994, Conn et al., 1997], quadratic [Powell, 2002], or polynomial [Shashaani et al., 2018] interpolation. The practical potential of RBF interpolation within the trust-region framework was investigated by Wild et al. [2008]. With regard to theory, fast convergence rates can be derived when the deriva-

tive of f is observable [Shi and Guo, 2008], but in the derivative-free setting, the main focus has been on global convergence to first-order [Wild and Shoemaker, 2011] and sometimes second-order [Conn et al., 2009b] critical points.

By contrast, methods based on Gaussian process regression typically do not explicitly distinguish between local and global search, but rather accomplish this tradeoff through a stochastic metamodel with built-in uncertainty quantification. The most popular algorithmic approach in this stream is expected improvement [Jones et al., 1998] and its many variants [Sasena et al., 2002, Huang et al., 2006]. The theory primarily focused on the pointwise consistency of the metamodel [Vazquez and Bect, 2010a] until the convergence rate analysis of Bull [2011], which was discussed earlier. Closely related is the probability of improvement criterion [Zhigljavsky and Zilinskas, 2008], which motivated the rate analysis of Calvin et al. [2018], also discussed previously.

Lastly, RBF interpolation has had a long history in numerical analysis [Buhmann, 2003] outside the setting of global optimization. Our analysis draws on this literature, specifically theory by Wu and Schaback [1993] characterizing the convergence rate of the estimation error (under RBF interpolation) given an arbitrary collection of design points. The first RBF-based global optimization procedure was proposed by Gutmann [2001], with later improvements by Regis and Shoemaker [2007] and Holmström [2008]. Sampling in these papers is based on a measure of the smoothness of the interpolation, with additional logic for balancing global and local search. Other sampling criteria have also been considered: for example, the method of Regis and Shoemaker [2005] aims to spread out the design points to avoid excessive clustering. Extensions include parallelized methods [Regis and Shoemaker, 2009] and hybrid methods combining RBF with ideas from

coordinate search [Regis and Shoemaker, 2013]. Much of this work is computationally oriented and focuses on complex engineering applications, with the theory mostly limited to global convergence. In the computer science community, Srinivas et al. [2010] derived rate results that apply to RBFs, but the setting there is online learning, where one optimizes cumulative error over time, rather than the offline setting more typical of global optimization (global search plays a much greater role in offline algorithms).

Overall, the approach and results presented here are intended, not to supplant the existing work on global optimization with RBFs, but to complement it from a theoretical viewpoint. Our paper adds to a very small number of prior studies of convergence rates for derivative-free optimization.

2.3 Problem statement and RBF interpolation

Let f be a function defined on a compact and connected domain $\mathcal{X} \subseteq \mathbb{R}^d$, and suppose that $x^* = \arg\min_{x \in \mathcal{X}} f(x)$ is the unique global minimizer of f. Let $\{x_n\}_{n=1}^N$ be a finite sequence of *design points* in $\mathbb{R}^d$, where for each x_n we observe $f(x_n)$ without noise. The design points can be pre-determined by the decision-maker or chosen adaptively; for the moment, however, suppose that they are simply given and that all the observations have been made. Using these observations, we construct a radial basis function (RBF) interpolation $\hat{f}_N$ of f and use $\hat{x}_N^* = \arg\min_{x \in \mathcal{X}} \hat{f}_N(x)$ as our estimate of x^*.

The RBF interpolation follows the method in Wu and Schaback [1993]. In order to apply the result in Wu and Schaback [1993] we adopt their assumptions which are introduced here. Let $\mathcal{P}_q$ be a space of polynomial functions on $\mathbb{R}^d$ with total order not

exceeding q; when $q > 0$, suppose that, if for some $p \in \mathcal{P}_q$ we have $p(x_n) = 0$ for all $n = 1, ..., N$, then $p \equiv 0$. Now let $\phi : \mathbb{R}_+ \to \mathbb{R}$ be a function chosen to make the mapping $r \mapsto \phi(\sqrt{r})$ conditionally positive definite of order q. This conditional positive definiteness means that for all N the $N \times N$ kernel matrix, whose (i, j)th entry is $\phi(\|x_i - x_j\|)$, is positive definite on the set of $u \in \mathbb{R}^N$ satisfying

$$\sum_{n=1}^{N} u_n p_i(x_n) = 0, \qquad i = 1, ..., Q,$$

where u_n is the n-th entry of u, $Q = \binom{q+d-1}{d}$ is the dimension of $\mathcal{P}_q$, and $(p_1, ..., p_N)$ is any basis of $\mathcal{P}_q$. The interpolation $\hat{f}_N$ then has the form

$$\hat{f}_N(x) = \sum_{n=1}^{N} a_n \phi(\|x - x_n\|) + \sum_{i=1}^{Q} a'_i p_i(x),$$

where the coefficients $(a_1, ..., a_N)$ and $\left(a'_1, ..., a'_Q\right)$ constitute the solution to the linear system

$$\begin{aligned} \sum_{n=1}^{N} a_n \phi(\|x_j - x_n\|) + \sum_{i=1}^{Q} a'_i p_i(x_j) &= f(x_j), \qquad j = 1, ..., N, \\ \sum_{n=1}^{N} a_n p_i(x_n) &= 0, \qquad i = 1, ..., Q. \end{aligned}$$

The conditional positive definiteness assumptions on ϕ guarantee that this system is non-singular.

Given the RBF ϕ, let $\psi(x) = \phi(\|x\|)$ and take $\tilde{\psi}$ to be the Fourier transform of ψ.

We define

$$c_{f,\phi}^2 = \int_{\mathbb{R}^d} \left|\tilde{f}(x)\right|^2 \tilde{\psi}(x)^{-1} dx$$

and require f to satisfy $c_{f,\phi}^2 < \infty$, a condition that is also imposed in Wu and Schaback [1993]. Moreover, although f itself is not required to be convex, for our proof we assume local strong convexity of f around its global minimizer.

To measure the local density of the design points, denote by

$$h_{\mathcal{D}} = \sup_{x \in \mathcal{D}} \inf_{n=1,\ldots,N} \|x - x_n\|_2 \tag{2.1}$$

the *mesh norm* of an arbitrary compact subset $\mathcal{D}$. The naming is adopted following the community of interpolation. Letting $\mu \in \mathbb{N}^d$ with $|\mu| := \sum_j \mu_j$, we use the standard multi-index notation $f^{(\mu)}$ for the function obtained after sequentially applying to f the μ_jth-order partial derivative with respect to x_j, $j = 1, \ldots, d$. Also let $B(x, r) = \{y : \|x - y\| \leq r\}$ be the ball of radius $r > 0$ centered at $x \in \mathbb{R}^d$. We now formally state the assumptions we make.

Assumption 2.1. *Assume that $\mathcal{X}$ is compact and connected, and f has unique global minimizer. Let the kernel function ϕ satisfy the (conditional) positive definiteness introduced above, which is necessary to cite the result of [Wu and Schaback, 1993]. Suppose that f is C^2 on $\mathcal{X}$ with $c_{f,\phi}^2 < \infty$, and the RBF ϕ is C^2 on $(0, \infty)$ and C^4 in a neighborhood of zero, with $k \geq 2$. Let $k = \frac{1}{2}s_\infty$ where s_∞ satisfies*

$$0 < \tilde{\psi}(t) \leq c_\psi \|t\|^{-d-s_\infty}$$

for $\|t\| \to \infty$. *Lastly, there exists a closed ball centered at* x^* *with strictly positive radius where* f *is strongly convex.*

This condition on s_∞ is satisfied by many commonly used kernels. For example, if $\phi(r) = e^{-\alpha r^2}$ is the Gaussian kernel (for some $\alpha > 0$), we may have arbitrarily large values of s_∞, leading to arbitrarily large values of k, which in effect causes $\hat{x}_n^*$ to converge to x^* even more quickly than the rate we derive in Section 3.3.4. However, k will be bounded for other types of kernels.

With these preliminaries, we can now state a result from Wu and Schaback [1993] that will be referenced and applied throughout this chapter.

Lemma 2.1. *[Wu and Schaback, 1993] With assumption 2.1 made, given* $\rho > 0$, *there exists* $k \in \mathbb{N}_+$ *and* $C \in \mathbb{R}_+$ *such that, for any* $\{x_n\}_{n=1}^N$, *any* N *and any* $x \in \mathcal{X}$ *satisfying* $h_{B(x,\rho)\cap\mathcal{X}} < h_0$, *with* h_0 *being a constant whose value depends on* k, *we have the inequality*

$$\left|\hat{f}_N^{(\mu)}(x) - f^{(\mu)}(x)\right| \leq c_{f,\phi} C h_{B(x,\rho)\cap\mathcal{X}}^{k-|\mu|}$$

for all μ *satisfying* $|\mu| \leq k$.

2.4 Properties of local mesh norms

Below, we show the equivalence of several measures of local data density, including the basic local mesh norm defined in (2.1), on a particular class of domains. The relationship between these measures will be useful in the subsequent analysis as we will draw on results obtained for different mesh norms by different research communities.

Definition 2.1. *A compact set $\mathcal{D} \subseteq \mathbb{R}^d$ is shape-regular if there exists a continuously differentiable bijection $L_{\mathcal{D}}$, mapping points from either $[0,1]^d$ or $B(0,1)$ (either can be chosen as the domain) onto $\mathcal{D}$, whose Jacobian has nonzero determinant everywhere on the domain.*

Definition 2.2. *Let $\mathcal{X}_N = \{x_n\}_{n=1}^N \subsetneq \mathcal{X}$ and define*

$$\begin{aligned}
\breve{h}_{\mathcal{D}}(\mathcal{X}_N) &= \sup_{y\in\mathcal{D}} \inf_{x\in\mathcal{X}_N\cap\mathcal{D}} \|x-y\|_2, \\
\bar{h}_{\mathcal{D}}(\mathcal{X}_N) &= \sup_{y\in\mathcal{D}} \inf_{x\in\mathcal{X}_N\cup\partial\mathcal{D}} \|x-y\|_2, \\
\tilde{h}_{\mathcal{D}}(\mathcal{X}_N) &= \sup_{y\in\mathcal{D}} \inf_{x\in\mathcal{X}_N\cup\partial\mathcal{D}} \|x-y\|_\infty,
\end{aligned}$$

when $\mathcal{X}_N \cap \mathcal{D} \neq \emptyset$. For simplicity, we may omit the explicit dependence of these quantities on $\mathcal{X}_N$ from the notation when there is no ambiguity. Note that they are not really norms.

To compare these and other quantities, we introduce the following notation. For two positive sequences $\{F_n^1\}_{n=1}^\infty$ and $\{F_n^2\}_{n=1}^\infty$ in $\mathbb{R}$, we write $F_n^1 \lesssim F_n^2$ if there exists a constant c, independent of n, such that $F_n^1 \leq cF_n^2$ for all n. We write $F_n^1 \cong F_n^2$ when $\lim_{n\to\infty} \frac{F_n^1}{F_n^2} = 1$ (note that this is stronger than having both $F_n^1 \lesssim F_n^2$ and $F_n^2 \lesssim F_n^1$).

Lemma 2.2. *Let $\mathcal{D} \subseteq \mathcal{X}$. For any $\{\mathcal{X}_N\}$, we have $\bar{h}_{\mathcal{D}} \lesssim \tilde{h}_{\mathcal{D}}$ and $\tilde{h}_{\mathcal{D}} \lesssim \bar{h}_{\mathcal{D}}$ (the above definition holds for $N \to \infty$). We also have $\bar{h}_{\mathcal{D}} \leq h_{\mathcal{D}} \leq \breve{h}_{\mathcal{D}}$.*

Proof: The first statement follows from the equivalence of norms in finite-dimensional spaces. The second statement follows from the relationship $\mathcal{X}_N \cap \mathcal{D} \subseteq \mathcal{X}_N \subseteq \mathcal{X} \cup \partial\mathcal{D}$. □

From Lemma 2.1, we know that on $\mathcal{X}$ the interpolation error is bounded by some power of the local mesh norm. Therefore, in order to study the convergence rate of the

interpolation error under any sampling scheme, we essentially require an appropriate decreasing rate for the local mesh norm. Janson [1987] derived such a rate for the specific case where the design points are sampled from a uniform distribution on $\mathcal{X}$. For convenience, we state this result here.

Lemma 2.3. *[Janson, 1987] Suppose that $\mathcal{X} = [0,1]^d$ or $\mathcal{X} = B(0,1)$. Suppose also that the design points $\{x_n\}_{n=1}^N$ are sampled i.i.d. from a uniform distribution on $\mathcal{X}$. Then,*

$$\tilde{h}_{\mathcal{X}} = O\left(\left(\frac{\log N}{N}\right)^{\frac{1}{d}}\right) \tag{2.2}$$

almost surely as $N \to \infty$. Furthermore, the multiplicative constant in (2.2) is nonrandom, i.e., the limit superior $\limsup_{N\to\infty} \tilde{h}_{\mathcal{X}} \left(\frac{\log N}{N}\right)^{-\frac{1}{d}}$ is a.s. equal to a deterministic quantity.

We also prove an analogous result for a more general case where the design points are sampled from an arbitrary probability distribution. This result helps motivate our subsequent focus on the epsilon-greedy policy, which uses uniform sampling for global search, because the convergence rate of the mesh norm is faster when the essential infimum of the sampling density is higher.

Lemma 2.4. *Suppose that $\mathcal{X} = [0,1]^d$ or $\mathcal{X} = B(0,1)$. Suppose also that the design points $\{x_n\}_{n=1}^N$ are sampled independently from a probability distribution with density $g : \mathcal{X} \to \mathbb{R}$. Let $g_{\min} = \operatorname{ess\,inf}_{x\in\mathcal{X}} g(x)$ and suppose that $g_{\min} > 0$. Then,*

$$\tilde{h}_{\mathcal{X}} = O\left(\left(\frac{\log(g_{\min} N)}{g_{\min} N}\right)^{\frac{1}{d}}\right)$$

almost surely.

Proof: We prove this lemma for $\mathcal{X} = [0,1]^d$, as the proof for the closed ball is similar. Define random variables $z_n \sim Bernoulli\,(g_{\min})$. Then, the distribution of the design points can be rewritten as follows: if $z_n = 1$, draw $x_n \sim U\left([0,1]^d\right)$, and if $z_n = 0$, draw x_n from a distribution with density $x \mapsto \frac{g(x)-g_{\min}}{1-g_{\min}}$. We can also denote by $\mathcal{X}_N^U = \{x_n : z_n = 1\}$ the subset of the design points coming from the uniform density, with $N^U = \sum_{n=1}^{N} z_n$ being the size of this subset. By the strong law of large numbers, $N^U \cong g_{\min} N$ almost surely.

Now, observe that for general $\mathcal{D} \subseteq \mathcal{X}$ and $\mathcal{Y}_1 \subseteq \mathcal{Y}_2 \subseteq \mathcal{X}$, we have $\tilde{h}_{\mathcal{D}}(\mathcal{Y}_1) \geq \tilde{h}_{\mathcal{D}}(\mathcal{Y}_2)$. Therefore, we have

$$\begin{aligned}
\tilde{h}_{[0,1]^d}(\mathcal{X}_N) &\leq \tilde{h}_{[0,1]^d}\left(\mathcal{X}_N^U\right) \\
&= O\left(\left(\frac{\log N^U}{N^U}\right)^{\frac{1}{d}}\right) \qquad \text{a.s.} \\
&= O\left(\left(\frac{\log\left(g_{\min} N\right)}{g_{\min} N}\right)^{\frac{1}{d}}\right) \qquad \text{a.s.},
\end{aligned}$$

as required. □

Using the above results, we can now obtain similar decreasing rates for other types of mesh norms defined on a more general domain.

Lemma 2.5. *Let $\mathcal{X}$ be compact and shape-regular, and suppose that the design points $x_1, ..., x_N$ are sampled independently from a uniform distribution on $\mathcal{X}$. Then,*

$$h_{\mathcal{X}}(\mathcal{X}_N) = O\left(\left(\frac{\log N}{N}\right)^{\frac{1}{d}}\right) \qquad \textit{a.s.}$$

Proof: By Definition 2.1, there exists a continuously differentiable bijection $L_{\mathcal{X}} : [0,1]^d \to \mathcal{X}$ whose Jacobian has nonzero determinant everywhere on the domain. Here we use $[0,1]^d$ as the domain of $L_{\mathcal{X}}$, but the proof is similar if $B(0,1)$ is used instead.

Since $L_{\mathcal{X}}$ is continuously differentiable on a compact set, it is Lipschitz. Then,

$$\begin{aligned} h_{\mathcal{X}} &\leq \breve{h}_{\mathcal{X}} \\ &= \sup_{y \in \mathcal{X}} \inf_{x \in \mathcal{X}} \|x - y\|_2 \\ &= \sup_{y \in [0,1]^d} \inf_{x \in L_{\mathcal{X}}^{-1}(\mathcal{X}_N)} \|L_{\mathcal{X}}(x) - L_{\mathcal{X}}(y)\|_2 \\ &\leq \sup_{y \in [0,1]^d} \inf_{x \in L_{\mathcal{X}}^{-1}(\mathcal{X}_N)} c_{L_{\mathcal{X}}} \|x - y\|_2, \end{aligned}$$

where $c_{L_{\mathcal{X}}}$ is the Lipschitz constant of $L_{\mathcal{X}}$.

Now, let us view $L_{\mathcal{X}}^{-1}(\mathcal{X}_N)$ as a set of design points on $[0,1]^d$. It can be easily shown that

$$\breve{h}_{[0,1]^d}\left(L_{\mathcal{X}}^{-1}(\mathcal{X}_N)\right) \leq \left(1 + \sqrt{d}\right) \bar{h}_{[0,1]^d}\left(L_{\mathcal{X}}^{-1}(\mathcal{X}_N)\right)$$

due to the geometry of the unit cube and the fact that $\bar{h}_{[0,1]^d}$ is the radius of the largest ball inside $[0,1]^d$ with no design points in its interior. By Lemma 2.2, we have

$$\begin{aligned} h_{\mathcal{X}} &\leq \sup_{y \in [0,1]^d} \inf_{x \in L_{\mathcal{X}}^{-1}(\mathcal{X}_N) \cup \partial\left([0,1]^d\right)} c_{L_{\mathcal{X}}} \left(1 + \sqrt{d}\right) \|x - y\|_2 \\ &\lesssim \sup_{y \in [0,1]^d} \inf_{x \in L_{\mathcal{X}}^{-1}(\mathcal{X}_N) \cup \partial\left([0,1]^d\right)} \|x - y\|_\infty \\ &= \tilde{h}_{[0,1]^d}\left(L_{\mathcal{X}}^{-1}(\mathcal{X}_N)\right). \end{aligned}$$

Let $|\mathcal{X}|$ be the volume of the domain $\mathcal{X}$ under Lebesgue measure. The design points

$L_{\mathcal{X}}^{-1}(\mathcal{X}_N)$ are drawn from a distribution with density

$$g_0(y) = \frac{1}{|\mathcal{X}|}\left|\det\left(\frac{dL_{\mathcal{X}}(y)}{dy}\right)\right|$$

whose essential infimum satisfies

$$g_{0,\min} = \operatorname*{ess\,inf}_{y\in[0,1]^d} g_0(y) > 0$$

due to the assumptions on $L_{\mathcal{X}}$. Then, by Lemma 2.4, for the mesh norm on $[0,1]^d$ with design points $L_{\mathcal{X}}^{-1}(\mathcal{X}_N)$, we have

$$\begin{aligned}\tilde{h}_{[0,1]^d}\left(L_{\mathcal{X}}^{-1}(\mathcal{X}_N)\right) &= O\left(\left(\frac{\log(g_{0,\min}N)}{g_{0,\min}N}\right)^{\frac{1}{d}}\right) \quad \text{a.s.}\\ &= O\left(\left(\frac{\log N}{N}\right)^{\frac{1}{d}}\right) \quad \text{a.s.}\end{aligned}$$

Putting everything together, we have

$$h_{\mathcal{X}}(\mathcal{X}_N) \lesssim \tilde{h}_{[0,1]^d}\left(L_{\mathcal{X}}^{-1}(\mathcal{X}_N)\right) = O\left(\left(\frac{\log N}{N}\right)^{\frac{1}{d}}\right)$$

almost surely, as required. □

2.5 Algorithm and main results

We now give a formal statement of the sequential algorithm used to select design points, and state our main theoretical results on its convergence rate. The remainder of

the chapter will consist of the proofs of these results.

Let

$$S(x,r) = \left\{ y \,:\, \|x-y\|_\infty \leq \frac{r}{2} \right\}$$

be the hypercube centered at x with sides of length r parallel to the coordinate axes. The ε-greedy algorithm, as defined in this paper, will randomly choose between uniform sampling inside this hypercube centered at the current-best solution $\hat{x}_n^*$ (local search) and uniform sampling on $\mathcal{X}$ (global search). Formally, let z_n be a Bernoulli random variable with success probability $b \cdot |\mathcal{X}|$, where $b \in \left(0, \frac{1}{|\mathcal{X}|}\right)$ is a constant and $|\mathcal{X}|$ is, again, the volume of $\mathcal{X}$ under Lebesgue measure. The success probability $b \cdot |\mathcal{X}|$ corresponds to $1-\varepsilon$ in the ε-greedy policy (we have $b \cdot |\mathcal{X}|\, 1-\varepsilon$), and governs the frequency of global search.

The distribution of the nth design point x_n is determined adaptively, after $x_0, ..., x_{n-1}$ and $f(x_0), ..., f(x_{n-1})$ have been observed, in the following way: when $z_n = 1$, x_n is sampled from the uniform distribution on $\mathcal{X}$. When $z_n = 0$, x_n is sampled from the uniform distribution on $S\left(\hat{x}_{n-1}^*, r_n\right)$, where the side length r_n will be discussed later, but is assumed small enough (or n large enough) to make $S\left(\hat{x}_{n-1}^*, r_n\right) \subseteq \mathcal{X}$. Treating z_n as a latent variable, x_n follows a distribution whose density g_n is a weighted average of two uniform distributions defined on distinct regions:

$$g_n(x) = \begin{cases} b, & x \in \mathcal{X} \setminus S\left(\hat{x}_{n-1}^*, r_n\right), \\ t_n & x \in S\left(\hat{x}_{n-1}^*, r_n\right), \end{cases} \tag{2.3}$$

where t_n is a constant satisfying $t_n > \frac{1}{|\mathcal{X}|} > b$, whose value can be determined by the

normalization condition

$$(t_n - b)\, r_n^d + b\,|\mathcal{X}| = 1 \tag{2.4}$$

from the values of b and r_n. Note that $\operatorname{ess\,inf}_{x\in\mathcal{X}} g_n(x) = b$ for all n.

In our analysis, the sequence $\{r_n\}$ of side lengths for the local search region is chosen according to

$$r_n = c_r \left(\frac{\log(bn)}{bn}\right)^{\frac{k}{2d}}, \tag{2.5}$$

where c_r is a large enough constant, and k is the constant in Lemma 2.1. With r_n chosen in this way, we obtain the following results. First, we bound the convergence rate of $\|\hat{x}_N^* - x^*\|$ on almost every sample path.

Theorem 2.6. *Assume that $\mathcal{X}$ is shape-regular, f is locally strongly convex on $B(x^*, \rho_0)$ for some constant radius $\rho_0 > 0$, the RBF parameter $k > 2$, and $\{r_n\}$ is chosen according to (2.5). Then, under assumption 2.1,*

$$\|\hat{x}_N^* - x^*\| = O\left(\left(\frac{\log N}{N}\right)^{\frac{k}{2d}} \left(\frac{\log(bN)}{bN}\right)^{\frac{k^2}{4d}}\right) \qquad \textit{a.s.} \tag{2.6}$$

It is worth noting that the decreasing rate of $\{r_n\}$ leads to an improvement in the convergence rate, in the form of the second factor in the right-hand side of (2.6). If the size of the local search region is constant, i.e., $r_n \equiv r_0$ for some $r_0 > 0$, (2.6) becomes

$$\|\hat{x}_N^* - x^*\| = O\left(\left(\frac{\log N}{N}\right)^{\frac{k}{2d}}\right).$$

Thus, the local search region should shrink over time.

The specific choice for the decreasing rate of $\{r_n\}$ is in some sense the best possible under the foundation provided by Lemma 2.1. Since the convergence rate of the local mesh norm (Lemma 2.3) is only able to guarantee a rate of $O\left(\left(\frac{\log N}{N}\right)^{\frac{1}{d}}\right)$, we are not able to shrink the radius of the local search region more quickly than this. At the same time, our bound will be loosened (as will be seen in the proof) if the radius vanishes more slowly than (2.5).

The second major result is a concentration inequality for the rate in Theorem 2.6. Essentially, for any given N there may be a non-negligible set of sample paths on which the rate in (2.6) does not hold. The probability measure of this set is bounded as follows.

Theorem 2.7. *Under the assumptions made in Theorem 2.6, for all large enough N,*

$$P\left(\|\hat{x}_N^* - x^*\| > c\left(\frac{\log N}{N}\right)^{\frac{k}{2d}}\left(\frac{\log(bN)}{bN}\right)^{\frac{k^2}{4d}}\right) \leq \frac{c'}{N},$$

where c, c' are constants.

We also considered a variant of the algorithm where b was also allowed to vary over n (recall from (2.3) that b is the value of the sampling density outside the local search region), as well as over $\mathcal{X}$. However, using non-uniform global search does not help the bound on the convergence rate, because our analysis relies on a connection between the estimation error and the mesh norm, and the convergence rate of the mesh norm for an arbitrary density g depends on $\operatorname{ess\,inf}_x g(x)$ as was seen in Lemma 2.4. If we then use uniform global search, but allow b to vary over time, we find that the bound becomes worse when $\liminf_n b_n = 0$ (in fact, if b_n vanishes too quickly, we may not even have $\hat{x}_N^* \to x^*$). On the other hand, when b_n varies between constant, strictly positive lower

and upper bounds, the order of the rate does not change, and these bounds only contribute to the multiplicative constant. For this reason, we decided not to overcomplicate the presentation with these details, and have simply used a constant b in the following.

2.6 Proof of Theorem 2.6

The proof of this result is separated into two parts. Section 2.6.1 discusses situations where $\hat{x}_N^* \in \mathcal{X} \setminus B(x^*, \rho_0)$, while Section 2.6.2 covers cases where $\hat{x}_N^* \in B(x^*, \rho_0)$. This distinction is made because, as N increases, if $\hat{x}_N^*$ is inside $B(x^*, \rho_0)$, the local strong convexity of f makes the convergence behavior of the estimated optimal solution more tractable. Thus, our first task is to show that $\hat{x}_N^* \in B(x^*, \rho_0)$ a.s. for large enough N.

2.6.1 Convergence of $\hat{x}_N^*$ to a neighborhood of x^*

We first prove a technical lemma giving a lower bound for the probability that $\hat{x}_N^* \in B(x^*, \rho_0)$. Similarly to the proof of Lemma 2.4, we rewrite the distribution of x_n as follows. For any n, let $z_n \sim Bernoulli(b \cdot |\mathcal{X}|)$ be a latent variable; then, if $z_n = 1$, draw $x_n \sim U(\mathcal{X})$, and if $z_n = 0$, draw x_n from a distribution with density $x \mapsto \frac{g_n - b}{1 - b|\mathcal{X}|}$, where g_n is as in (2.3). We then denote by $\mathcal{X}_N^U = \{x_n : z_n = 1\}$ the subset of the design points coming from the uniform density, with $N^U = \sum_{n=1}^{N} z_n$ being the number of such samples.

Our analysis proceeds by deriving a lower bound on $P(\hat{x}_N^* \in B(x^*, \rho_0))$. This bound will eventually be shown to converge to 1. The first step is given in the following lemma.

Lemma 2.8. *Define* $\tilde{x} = \arg\min_{x \in \mathrm{cl}(\mathcal{X} \setminus B(x^*, \rho_0))} f(x)$ *and*

$$\mathcal{D} = \left\{ x \in \mathcal{X} \,:\, f(x) < \frac{1}{2}\left(f(\tilde{x}) + f(x^*)\right) \right\}.$$

There exists a positive constant c_w*, independent of the sampling policy and the design points, such that*

$$P\left(\hat{x}_N^* \in B(x^*, \rho_0)\right) \geq P\left(h_{\mathcal{X}}\left(\mathcal{X}_N^U\right) < c_w\right) + P\left(\bigcup_{n=1}^{N} \{x_n \in \mathcal{X}_N \cap \mathcal{D}\}\right) - 1. \quad (2.7)$$

The proof is moved to section 2.8 due to space considerations.

The next lemma, whose proof is also moved to section 2.8, bounds the first term on the right-hand side of (2.7). This bound is then used in the next result (Lemma 2.10) to further bound the left-hand side of (2.7).

Lemma 2.9. *Let* c_w *be the constant obtained from Lemma 2.8. There exists another positive constant* $\bar{c}_w$ *such that*

$$P\left(h_{\mathcal{X}}\left(\mathcal{X}_N^U\right) < c_w\right) \geq 1 - \bar{c}_w e^{-b^2 |\mathcal{X}|^2 N/2}.$$

Lemma 2.10. *There exists a positive constant* $\hat{c}_w$ *such that*

$$P\left(\hat{x}_N^* \in \mathcal{X} \setminus B(x^*, \rho_0)\right) \lesssim e^{-\hat{c}_w N}.$$

Proof: Let $\tilde{x}$ and $\mathcal{D}$ be as in the statement of Lemma 2.8. We calculate

$$\begin{aligned} P\left(\bigcup_{n=1}^{N}\left\{x_n \in \mathcal{X}_N \cap \mathcal{D}\right\} \mid N^U\right) &\geq 1 - P\left(\mathcal{X}_N^U \subseteq \mathcal{X} \setminus \mathcal{D} \mid N^U\right) \\ &= 1 - \left(1 - \frac{|\mathcal{D}|}{|\mathcal{X}|}\right)^{N^U}. \end{aligned}$$

Taking the expectation over the distribution of N^U, we obtain

$$1 - P\left(\bigcup_{n=1}^{N}\left\{x_n \in \mathcal{X}_N \cap \mathcal{D}\right\}\right) = \mathbb{E}\left(\left(1 - \frac{|\mathcal{D}|}{|\mathcal{X}|}\right)^{N^U}\right).$$

By the independence of $\{z_n\}$, we obtain

$$\begin{aligned} \mathbb{E}\left(\left(1 - \frac{|\mathcal{D}|}{|\mathcal{X}|}\right)^{N^U}\right) &= \mathbb{E}\prod_{n=1}^{N}\left(1 - \frac{|\mathcal{D}|}{|\mathcal{X}|}\right)^{z_n} \\ &= \left(\left(1 - \frac{|\mathcal{D}|}{|\mathcal{X}|}\right) b\,|\mathcal{X}| + 1 - b\,|\mathcal{X}|\right)^{N}. \end{aligned}$$

Then,

$$1 - P\left(\bigcup_{n=1}^{N}\left\{x_n \in \mathcal{X}_N \cap \mathcal{D}\right\}\right) = e^{N\log(1-b|\mathcal{D}|)}. \tag{2.8}$$

Combining (2.8) with Lemmas 2.8 and 2.9, we obtain

$$\begin{aligned} P\left(\hat{x}_N^* \in B\left(x^*, \rho_0\right)\right) &\geq P\left(h_{\mathcal{X}}\left(\mathcal{X}_N^U\right) < c_w\right) + P\left(\bigcup_{n=1}^{N}\left\{x_n \in \mathcal{X}_N \cap \mathcal{D}\right\}\right) - 1 \\ &\geq 1 - \bar{c}_w e^{-b^2|\mathcal{X}|^2 N/2} - e^{N\log(1-b|\mathcal{D}|)}, \end{aligned}$$

whence the desired result follows. □

From Lemma 2.10, it follows that

$$\sum_{N=1}^{\infty} P\left(\hat{x}_N^* \notin B\left(x^*, \rho_0\right)\right) < \infty.$$

By a direct application of the Borel-Cantelli lemma, we find that

$$P\left(\limsup_{N\to\infty}\left\{\hat{x}_N^* \notin B\left(x^*, \rho_0\right)\right\}\right) = 0,$$

which means that, asymptotically, $\hat{x}_N^* \in B\left(x^*, \rho_0\right)$ w.p. 1.

2.6.2 Convergence rate around x^*

Using the results of Section 2.6.1, we know that there is an almost surely finite random number N_ω so that $\hat{x}_N^* \in B\left(x^*, \rho_0\right)$ whenever $N > N_\omega$. This condition will be occasionally made in this section only. Keeping the notation introduced previously, we begin by applying the strong law of large numbers to N^U. Since

$$\sum_{n=1}^{\infty} \frac{1}{n^2} Var\left(z_n\right) < \infty,$$

we have $\frac{N^U - \mathbb{E}\left(N^U\right)}{N} \to 0$ as $N \to \infty$, i.e., $N^U \cong |\mathcal{X}|\, bN$ a.s. Hence,

$$h_{\mathcal{X}}\left(\mathcal{X}_N^U\right) = O\left(\left(\frac{\log N^U}{N^U}\right)^{\frac{1}{d}}\right) = O\left(\left(\frac{\log\left(bN\right)}{bN}\right)^{\frac{1}{d}}\right) \qquad \text{a.s.}$$

Then,

$$h_{B(x^*,\rho_0)}\left(\mathcal{X}_N\right) \leq h_{\mathcal{X}}\left(\mathcal{X}_N\right) \leq h_{\mathcal{X}}\left(\mathcal{X}_N^U\right) = O\left(\left(\frac{\log\left(bN\right)}{bN}\right)^{\frac{1}{d}}\right) \tag{2.9}$$

almost surely as $N \to \infty$.

Applying Lemma 2.1 on $B\left(x^*, \rho_0\right)$, we have

$$\sup_{x \in B(x^*,\rho_0)} \left|\hat{f}_N^{(\mu)}\left(x\right) - f^{(\mu)}\left(x\right)\right| \leq c_{f,\phi} C h_{B(x^*,\rho_0)}^{k-|\mu|}$$

for $\mu \in \mathbb{N}^d$ with $|\mu| \leq k$. Because $k > 2$ by assumption, we can quantify the approximation error of $\hat{f}_N$ and its Hessian as

$$\sup_{x \in B(x^*,\rho_0)} \left|\hat{f}_N\left(x\right) - f\left(x\right)\right| \leq c_{f,\phi} C h_{B(x^*,\rho_0)}^{k}$$

and

$$\sup_{x \in B(x^*,\rho_0)} \left|\frac{\partial^2 \hat{f}_N\left(x\right)}{\partial x_i \partial x_j} - \frac{\partial^2 f\left(x\right)}{\partial x_i \partial x_j}\right| \leq c_{f,\phi} C h_{B(x^*,\rho_0)}^{k-2}, \qquad i,j = 1,...,d.$$

At the same time, letting $H_f\left(x\right)$ and $H_{\hat{f}_N}\left(x\right)$ be the Hessian matrices of (respectively) f and $\hat{f}_N$ at x, we can write the expansion

$$f\left(\hat{x}_N^*\right) = f\left(x^*\right) + \nabla f\left(x^*\right)^\top \left(\hat{x}_N^* - x^*\right) + \frac{1}{2}\left(\hat{x}_N^* - x^*\right)^\top H_f\left(\breve{x}\right)\left(\hat{x}_N^* - x^*\right)$$

for some $\breve{x}$ on the segment joining $\hat{x}_N^*$ and x^*, i.e.,

$$f\left(\hat{x}_N^*\right) - f\left(x^*\right) = \frac{1}{2}\left(\hat{x}_N^* - x^*\right)^\top H_f\left(\breve{x}\right)\left(\hat{x}_N^* - x^*\right). \tag{2.10}$$

Similarly, we have

$$\hat{f}_N(x^*) - \hat{f}_N(\hat{x}_N^*) = \frac{1}{2}(\hat{x}_N^* - x^*)^\top H_{\hat{f}_N}(\breve{x}_N)(\hat{x}_N^* - x^*) \tag{2.11}$$

for some $\breve{x}_N$ on the segment joining $\hat{x}_N^*$ and x^*. Adding (2.10) and (2.11), we obtain the upper bound

$$\begin{aligned}
& \frac{1}{2}(\hat{x}_N^* - x^*)^\top \left(H_f(\breve{x}) + H_{\hat{f}_N}(\breve{x}_N)\right)(\hat{x}_N^* - x^*) \\
= \; & f(\hat{x}_N^*) - \hat{f}_N(\hat{x}_N^*) + \hat{f}_N(x^*) - f(x^*) \\
\leq \; & \left|f(\hat{x}_N^*) - \hat{f}_N(\hat{x}_N^*)\right| + \left|\hat{f}_N(x^*) - f(x^*)\right| \\
\leq \; & 2c_{f,\phi} C h^k_{B(x^*,\rho_0)}(\mathcal{X}_N).
\end{aligned} \tag{2.12}$$

A lower bound can be obtained via

$$\begin{aligned}
& \frac{1}{2}(\hat{x}_N^* - x^*)^\top \left(H_f(\breve{x}) + H_{\hat{f}_N}(\breve{x}_N)\right)(\hat{x}_N^* - x^*) \\
= & \frac{1}{2}(\hat{x}_N^* - x^*)^\top \left(H_f(\breve{x}) + H_f(\breve{x}_N) - \left(H_f(\breve{x}_N) - H_{\hat{f}_N}(\breve{x}_N)\right)\right)(\hat{x}_N^* - x^*) \\
\geq & \frac{1}{2}\|\hat{x}_N^* - x^*\|^2 \left[\lambda_{\min}(H_f(\breve{x})) + \lambda_{\min}(H_f(\breve{x}_N)) - \lambda_{\max}\left(H_f(\breve{x}_N) - H_{\hat{f}_N}(\breve{x}_N)\right)\right] \\
\geq & \frac{1}{2}\|\hat{x}_N^* - x^*\|^2 \left[2 \inf_{x \in B(x^*,\rho_0)} \lambda_{\min}(H_f(x)) - \sup_{x \in B(x^*,\rho_0)} \lambda_{\max}\left(H_f(x) - H_{\hat{f}_N}(x)\right)\right].
\end{aligned} \tag{2.13}$$

Note that the constant

$$\lambda_0 = \inf_{x \in B(x^*,\rho_0)} \lambda_{\min}(H_f(x))$$

satisfies $\lambda_0 > 0$ by the assumption that f is locally strongly convex inside $B(x^*, \rho_0)$. A

further lower bound can be obtained from the following technical lemma (whose proof is deferred to section 2.8).

Lemma 2.11. *For a positive definite $d \times d$ matrix A satisfying $|A_{ij}| \leq t_A$ for all i, j, then $\lambda_{\max}(A) \leq d \cdot t_A$.*

Applying Lemma (2.11) to (2.13) leads to

$$\begin{aligned} & \frac{1}{2}\left(\hat{x}_N^* - x^*\right)^\top \left(H_f\left(\breve{x}\right) + H_{\hat{f}_N}\left(\breve{x}_N\right)\right)\left(\hat{x}_N^* - \hat{x}\right) \\ \geq\ & \frac{1}{2}\|\hat{x}_N^* - x^*\|^2 \left[2\lambda_0 - d \sup_{x \in B(x^*,\rho_0)} \max_{i,j} \left(H_f(x) - H_{\hat{f}_N}(x)\right)_{i,j}\right] \\ \geq\ & \frac{1}{2}\|\hat{x}_N^* - x^*\|^2 \left[2\lambda_0 - dc_{f,\phi} C h_{B(x^*,\rho_0)}^{k-2}\right]. \end{aligned} \tag{2.14}$$

Combining (2.12) with (2.14) yields

$$\|\hat{x}_N^* - x^*\| \leq \left(\frac{4c_{f,\phi} C h_{B(x^*,\rho_0)}^{k}\left(\mathcal{X}_N\right)}{2\lambda_0 - dc_{f,\phi} C h_{B(x^*,\rho_0)}^{k-2}}\right)^{\frac{1}{2}} = O\left(h_{B(x^*,\rho_0)}^{\frac{k}{2}}\left(\mathcal{X}_N\right)\right) \tag{2.15}$$

almost surely when $N > N_\omega$ and $h_{B(x^*,\rho_0)}$ is small enough. Considering the decreasing rate of the mesh norm obtained from Lemma 2.5, this in turn implies

$$\|\hat{x}_N^* - x^*\| = O\left(\left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}}\right) \qquad \text{a.s.} \tag{2.16}$$

The rate in (2.16) can be improved by narrowing the local search region over time, as long as both x^* and $\hat{x}_N^*$ are elements of each region in the sequence, and we collect infinitely many samples from these regions. The decay rate of $\{r_n\}$ begins to play an important role in ensuring that these conditions hold. If r_n decays too slowly, we will undersample

in the local regions; if r_n decays too quickly, our local regions may fail to cover x^*.

It turns out that the best possible rate for r_n is $c_r \left(\frac{\log(bn)}{bn}\right)^{\frac{k}{2d}}$ for some large enough and deterministic constant $c_r > 0$. With this specific choice, the following result is obtained.

Lemma 2.12. *Suppose that x_n is sampled from the density g_n defined in (2.3). Let $r_n = c_r \left(\frac{\log(bn)}{bn}\right)^{\frac{k}{2d}}$ for $c_r > 0$ and define*

$$S_n = S\left(x^*, 2c_b \left(\frac{\log(bn)}{bn}\right)^{\frac{k}{2d}}\right).$$

Then, for c_r large enough, there exist a deterministic constant $c_b > 0$ and a random integer $n_r > \max\{N_\omega, 1\}$ such that when $N > n_r$,

$$S_N \subseteq \bigcap_{n=n_r}^{N} S\left(\hat{x}^*_{n-1}, r_n\right)$$

almost surely.

Proof: We have already obtained (2.16). From this result, there exist a deterministic constant c_b and a random integer n_r such that, almost surely for all $n \geq n_r$,

$$\|\hat{x}^*_N - x^*\| \leq c_b \left(\frac{\log(bn)}{bn}\right)^{\frac{k}{2d}}$$

and (by the equivalence of vector norms),

$$\|\hat{x}^*_N - x^*\|_\infty \leq c_b \left(\frac{\log(bn)}{bn}\right)^{\frac{k}{2d}}.$$

Let c_r be large enough to satisfy that, uniformly for all values of n_r,

$$r_n > c_b \left(\frac{\log\left(b\left(n-1\right)\right)}{b\left(n-1\right)} \right)^{\frac{k}{2d}} + c_b \left(\frac{\log\left(bn\right)}{n} \right)^{\frac{k}{2d}}, \qquad \forall n \geq n_r.$$

Then, we almost surely have $r_n > \|\hat{x}^*_{n-1} - x^*\|_\infty + \|\hat{x}^*_n - x^*\|_\infty$, whence $x^* \in S\left(\hat{x}^*_{n-1}, r_n\right)$ and $S_n \subseteq S\left(\hat{x}^*_{n-1}, r_n\right)$. Similarly, for all n, N satisfying $N \geq n \geq n_r > \max\{N_\omega, 1\}$, we almost surely have $r_n > \|\hat{x}^*_{n-1} - x^*\|_\infty + \|\hat{x}^*_N - x^*\|_\infty$, whence we obtain $S_N \subseteq S\left(\hat{x}^*_{n-1}, r_n\right)$, as required. □

From Lemma 2.12, it follows that the ball

$$B_N = B\left(x^*, c_b \left(\frac{\log\left(bN\right)}{bN} \right)^{\frac{k}{2d}} \right)$$

inscribed in S_N satisfies $B_N \subseteq \bigcap_{n=n_r}^{N} S\left(\hat{x}^*_{n-1}, r_n\right)$. Additionally, from the results of Section 2.6.1 we have $\hat{x}^*_N \in B_N$ almost surely whenever $N \geq n_\omega$, a random number greater than n_r that is almost surely finite.

From (2.3), we almsot surely have $g_N\left(x\right) = t_N$ for all $x \in B_N$ and $N \geq n_\omega > n_r$. Let

$$\mathcal{X}^{B_N} = \{x_{n_\omega}, x_{n_\omega+1}, ..., x_N\} \cap B_N, \qquad N^{B_N} = \text{card}\left(\mathcal{X}^{B_N}\right).$$

Any $x \in \mathcal{X}^{B_N}$ was drawn from a uniform distribution on B_N, conditional on n_ω. The volume of B_N satisfies

$$|B_N| = c_B \cdot c_b^d \left(\frac{\log\left(bN\right)}{bN} \right)^{\frac{k}{2}}$$

for some constant c_B.

Conditional on n_ω, for any fixed $N \geq n_\omega$, let $z'_{n_\omega}|n_\omega, ..., z'_N|n_\omega$ be independent Bernoulli random variables, with the success probability of $z'_n|n_\omega$ being $t_n |B_N|$ for $n = n_\omega, ..., N$. Sampling x_n from the density g_n is equivalent to sampling uniformly on B_N if $z'_n|n_\omega = 1$, and sampling from a distribution with density function $x \mapsto g_n(x) - t_n 1_{B_N}(x)$ if $z'_n|n_\omega = 0$. Then, since $N^{B_N}|n_\omega = \sum_{n=n_\omega}^{N} z'_n|n_\omega$ it follows by the strong law of large numbers conditional on n_ω and the almost sure finiteness of n_ω that

$$\frac{N^{B_N}}{N - n_\omega + 1} \cong t_n |B_N| \qquad a.s.$$

Note that B_N is shape-regular because we can use a shifting and scaling mapping on $B(0,1)$ into B_N. The scaling factor is precisely $c_b \left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}}$. The mesh norm of a set $\mathcal{D}$ will be invariant if $\mathcal{D}$ is translated and the sampling density is changed accordingly. If $\mathcal{D}$ is scaled isotropically and the sampling density is changed accordingly, the mesh norm will be scaled by the same scaling factor.

Conditional on n_ω and N^{B_N}, $\mathcal{X}^{B_N}$ contains N^{B_N} points sampled uniformly on B^N. By Lemma 2.3, the equivalence of different types of mesh norms on a hypercube (Lemma 2.2), and the properties of the bijection L^{B_N} obtained from shape-regularity, we have

$$\begin{aligned} h_{B_N}\left(\mathcal{X}^{B_N}\right) &= O\left(c_b \left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}} \left(\frac{\log N^{B_N}}{N^{B_N}}\right)^{\frac{1}{d}}\right) \\ &= O\left(c_b \left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}} \left(\frac{\log\left(t_N \cdot |B_N| (N - n_\omega + 1)\right)}{t_N \cdot |B_N| (N - n_\omega + 1)}\right)^{\frac{1}{d}}\right) \\ &= O\left(\frac{\log(t_n \cdot |B_N| \cdot N)}{t_n N}\right), \end{aligned}$$

with each "inequality" holding a.s., as long as $t_N \cdot |B_N| \cdot N$ does not converge to zero.

We will now prove this last assertion.

Because $\hat{x}_N^*, x^* \in B_N$ when $N \geq n_\omega$, with $\|\hat{x}_N^* - x^*\| = O\left(\left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}}\right)$ almost surely, we can obtain the a.s. inequality

$$\|\hat{x}_N^* - x^*\|^2 \leq \frac{4c_{f,\phi}Ch_{B_N}^k\left(\mathcal{X}_N\right)}{2\lambda_N - dc_{f,\phi}Ch_{B_N}^{k-2}}, \tag{2.17}$$

similarly to (2.15), for h_{B_N} small enough and $N \geq n_\omega$, by repeating the steps of the analysis done for $B\left(x^*, \rho_0\right)$ using the strong convexity of f on B_N. The quantity λ_N in (2.17) is defined as $\lambda_N = \inf_{x \in B_N} \lambda_{\min}\left(H_f\left(x\right)\right)$.

Recall that $h_{B_N}\left(\mathcal{X}_N\right) \leq h_{B_N}\left(\mathcal{X}^{B_N}\right)$, and observe that $\lambda_N \to \lambda_{\min}\left(H_f\left(x^*\right)\right)$ as $N \to \infty$. It follows, analogously to (2.16), that

$$\|\hat{x}_N^* - x^*\| = O\left(h_{B_N}^{\frac{k}{2}}\left(\mathcal{X}_N\right)\right) = O\left(\left(\frac{\log\left(t_N \cdot |B_N| \cdot N\right)}{t_N N}\right)^{\frac{k}{2d}}\right) \qquad \text{a.s.} \tag{2.18}$$

With $r_N = c_r\left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}}$, the rate of t_N follows from (2.4), and is given by

$$\begin{aligned} t_N &= b + \frac{1 - b\left|\mathcal{X}\right|}{r_N^d} \\ &= b + \frac{1 - b\left|\mathcal{X}\right|}{c_r^d}\left(\frac{\log\left(bN\right)}{bN}\right)^{-\frac{k}{2}} \\ &= O\left(\left(\frac{\log\left(bN\right)}{bN}\right)^{-\frac{k}{2}}\right). \end{aligned} \tag{2.19}$$

From this it follows that

$$r_N^d t_N N = \left(1 - b\left|\mathcal{X}\right| + br_N^d\right) N = O\left(N\right),$$

which means that $t_N \cdot |B_N| \cdot N = O(N)$, as required. Finally, substituting (2.19) into (2.18) yields

$$\|\hat{x}_N^* - x^*\| = O\left(\left(\frac{\log N}{N}\right)^{\frac{k}{2d}} \left(\frac{\log(bN)}{bN}\right)^{\frac{k^2}{4d}}\right) \qquad \text{a.s.}$$

as claimed by Theorem 2.6.

Note that, from (2.18), we see that the bound will be tightened if t_N becomes larger. The tightest possible bound is obtained when r_N follows (2.5). At the same time, we cannot make r_N vanish more quickly than the rate obtained in (2.16) using the properties of the mesh norm.

2.7 Proof of Theorem 2.7

For readability, the proof of this result is separated into three parts:

- In Section 2.7.1, the main goal is to derive a concentration inequality for the mesh norm defined on $[0,1]^d$ under uniform sampling. This inequality is obtained from an analysis of uniform sampling applied to a finite partition of the domain, whose size then grows at a suitably chosen rate.

- In Section 2.7.2, we then extend the results of the first part to the general domain $\mathcal{X}$ under epsilon-greedy sampling.

- Finally, in Section 2.7.3, the concentration inequalities obtained for the mesh norm are converted into analogous results for the estimation error $\|\hat{x}_N^* - x^*\|$.

Each part consists of multiple technical steps grouped together to make each subsection as self-contained as possible.

2.7.1 Concentration of mesh norm on $[0,1]^d$ under uniform sampling

In this section, we derive a concentration inequality on the mesh norm $h_{[0,1]^d}\left(\mathcal{X}_0^U\right)$ with $\mathcal{X}_0^U$ being a set of uniformly sampled design points. To study the behaviour of the mesh norm in finite time, we partition the domain into a finite number of subsets, then grow the size of the partition as more design points are sampled at a suitably chosen rate. The growth rate of the partition size can be related to the declining behavior of the mesh norm.

Specifically, we consider $m = K^d$ sets $\zeta_{m,1}, ..., \zeta_{m,m}$ of the form $\left[\frac{i_1}{K}, \frac{i_1+1}{K}\right] \times \left[\frac{i_2}{K}, \frac{i_2+1}{K}\right] \times ... \times \left[\frac{i_d}{K}, \frac{i_d+1}{K}\right]$ with $i_j = 0, 1, ..., K-1$ for $j = 1, ...d$. We then have $\bigcup_i \zeta_{m,i} = [0,1]^d$ and $\zeta_{m,i} \cap \zeta_{m,j}$ has zero Lebesgue measure for any $i \neq j$ (also under the measure induced by the uniform distribution). Each subset $\zeta_{m,i}$ also has the same volume under either measure.

Given N design points, let M be the maximum number of subsets into which the domain can be partitioned (according to the method described above) such that each subset contains at least one design point in its interior. The probability that any design points will fall exactly on the joint boundary of two adjacent subsets is zero. Formally,

$$M = \max\left\{m \,:\, \forall i = 1, 2, ..., m,\ \mathcal{X}_N \cap \operatorname{int}\left(\zeta_{m,i}\right) \neq \emptyset\right\}.$$

Thus, M is a random variable that takes positive integer values and is dependent on N.

We are interested in the increasing rate of M as N becomes large.

First, for fixed positive integers m and N, let $W(N,m) = (W_1, ..., W_m)$ be a random vector following a multinomial distribution with parameters (N,m) and probability vector $\frac{1}{m} \cdot \delta$, where $\delta_j = 1$ for each component j. The following result (proved in section 2.8) calculates the probability that no subsets will be empty.

Lemma 2.13. *For any N and m,*

$$P(W(N,m) \succeq \delta) = \frac{1}{m^N} \sum_{j=0}^{m} \binom{m}{j} (-1)^{m-j} j^N. \tag{2.20}$$

Now observe that

$$P(M \geq m) = P(W(N,m) \succeq \delta),$$

because the maximum partition size is at least m if and only if, in the partition whose size is m, each subset contains at least one design point. In another way of writing, $P(M \geq m) = \frac{m!}{m^N} \left\{ {N \atop m} \right\}$, where $\left\{ {N \atop m} \right\}$ represents a Stirling number of the second kind, defined for general m, n as

$$\left\{ {n \atop m} \right\} = \frac{1}{m!} \sum_{j=0}^{m} \binom{m}{j} (-1)^{m-j} j^n.$$

By following the asymptotic analysis of Stirling numbers [Temme, 1993], we can obtain the following lemma for the increasing rate of M. The next result (whose proof is deferred to section 2.8) then derives a more useful form for this approximation.

Lemma 2.14. *[Temme, 1993] The Stirling number of the second kind* $\left\{ {n \atop m} \right\}$ *can be approximated (uniformly in* m*, and asymptotically as* $n \to \infty$*) as*

$$\left\{ {n \atop m} \right\} \cong e^T m^{n-m} \tau \left(\frac{n}{m} - 1 \right) \binom{n}{m}, \tag{2.21}$$

where

$$\begin{aligned} T &= \gamma(y_0) - n + m + (n-m) \log\left(\frac{n}{m} - 1 \right), \\ \tau(s) &= \sqrt{\frac{ms}{n(y_0 - s)}}, \\ \gamma(y) &= -n \log y + m \log\left(e^y - 1\right), \end{aligned}$$

and y_0 *is the solution of* $\frac{m}{n} y = 1 - e^{-y}$*.*

Lemma 2.15. *The Stirling number of the second kind* $\left\{ {n \atop m} \right\}$ *can be approximated (uniformly in* m*, and asymptotically as* $n \to \infty$*) as*

$$\left\{ {n \atop m} \right\} \cong e^{m-n} \left(\frac{\frac{n}{m} - 1}{\frac{n}{m} - w} \right)^{n-m} \frac{m^{n-m}}{w^m} \sqrt{\frac{\frac{n}{m} - 1}{\frac{n}{m}(1-w)}} \binom{n}{m},$$

where $w = -W_0\left(-\frac{n}{m} e^{-\frac{n}{m}}\right)$ *and* W_0 *is the upper branch of the Lambert* W *function.*

We can then apply Lemma 2.15 to obtain an asymptotic rate for $P(M \geq m)$ that is expressible in closed form as a function of N and m. Consequently, as N grows large, we can choose a suitable growth rate for m that would allow us to achieve a partition of size m w.p. 1. The following bound (proved in section 2.8) provides a sufficient condition for this growth rate to be suitable.

Lemma 2.16. *If $m \leq N$ and $m, N \to \infty$ with $\frac{N}{m} \to \infty$, then $P(M < m) \leq c_t' \frac{1}{N-m}$ for some constant $c_t' > 0$.*

With these technical results, we return to the decreasing rate of the mesh norm on $[0,1]^d$ under uniform sampling. The following result follows fairly straightforwardly from the preceding.

Lemma 2.17. *Let $\mathcal{X}_0^U$ be a set of N design points sampled independently from the uniform distribution on $[0,1]^d$. There exist constants $c_h, c_t > 0$ such that, for any $N > 1$,*

$$P\left(h_{[0,1]^d}\left(\mathcal{X}_0^U\right) > c_h\left(\frac{\log N}{N}\right)^{\frac{1}{d}}\right) \leq c_t \frac{1}{N}.$$

Proof: There exists a constant c_h' such that the statement $M \geq m$ (for any m) implies that $h_{[0,1]^d}\left(\mathcal{X}_0^U\right) \leq c_h' m^{-\frac{1}{d}}$. To match the almost sure convergence rate of the mesh norm, choose $m^* \cong \frac{N}{\log N}$. Then, there exist constants $c_h, c_t > 0$ such that, for any $N > 1$,

$$\begin{aligned} P\left(h_{[0,1]^d}\left(\mathcal{X}_0^U\right) > c_h\left(\frac{\log N}{N}\right)^{\frac{1}{d}}\right) &\leq P(M < m^*) \\ &\leq c_t' \frac{1}{N - m^*} \qquad (2.22) \\ &\leq c_t \frac{1}{N}, \end{aligned}$$

with (2.22) following from Lemma 2.16. □

2.7.2 Concentration of mesh norm on $\mathcal{X}$ under ε-greedy sampling

The final result of this section is an analog of Lemma 2.17 from the previous section with the general domain $\mathcal{X}$ and the design points $\mathcal{X}_N$ obtained from ε-greedy sampling.

We move to this more general case in several steps. First, we consider non-uniform sampling while keeping the domain as $[0,1]^d$.

Let $\mathcal{X}'_N = \{x'_1, ..., x'_N\}$ where each x'_n is sampled independently from a density g'_n with support $[0,1]^d$. Let $b'_n = \inf_{x\in[0,1]^d} g'_n(x)$ and suppose that $b'_n > 0$ and $\sum_{n=1}^N b'_n \to \infty$ as $N \to \infty$. Let z'_n be independent Bernoulli random variables with success probabilities b'_n so that, if $z'_n = 1$, then x'_n is sampled from a uniform distribution on $[0,1]^d$, and if $z'_n = 0$, then x'_n is sampled from the density $x \mapsto \frac{g'_n(x)-b'_n}{1-b'_n}$.

Let $N^{U'} = \sum_{n=1}^N z'_n$ and $\bar{b}'_N = \frac{1}{N}\sum_{n=1}^N b'_n$. A direct application of Hoeffding's inequality yields

$$\begin{aligned} P\left(\left|N^{U'} - \bar{b}'_N N\right| > \frac{1}{2}\bar{b}'_N N - 1\right) &\leq 2e^{-\frac{2}{N}\left(\frac{1}{2}\bar{b}'_N N - 1\right)^2} \\ &= o\left(\frac{1}{\bar{b}'_N N}\right). \end{aligned} \tag{2.23}$$

In later proofs, this will be combined with the following technical result (proved in section 2.8).

Lemma 2.18. *If* $\left|N^{U'} - \bar{b}'_N N\right| \leq \frac{1}{2}\bar{b}'_N N - 1$, *then*

$$\left(\frac{\log N^{U'}}{N^{U'}}\right)^{\frac{1}{d}} \leq 2^{\frac{1}{d}}\left(\frac{\log\left(\bar{b}'_N N\right)}{\bar{b}'_N N}\right)^{\frac{1}{d}}.$$

The next result is a concentration inequality for the mesh norm on $[0,1]^d$ under non-uniform sampling.

Lemma 2.19. *There exist constants $c_{h,1}, c_{t,1}$ such that, for large enough N,*

$$P\left(h_{[0,1]^d}\left(\mathcal{X}'_N\right) > c_{h,1}\left(\frac{\log\left(\bar{b}'_N N\right)}{\bar{b}'_N N}\right)^{\frac{1}{d}}\right) \leq c_{t,1}\frac{1}{\bar{b}'_N N}.$$

Proof: For notational compactness, denote the event

$$E = \left\{\left|N^{U'} - \bar{b}'_N N\right| \leq \frac{1}{2}\bar{b}'_N N - 1\right\}.$$

For any arbitrary positive value of $c_{h,1}$, we derive

$$\begin{aligned}
& P\left(h_{[0,1]^d}\left(\mathcal{X}'_N\right) > c_{h,1}\left(\frac{\log\left(\bar{b}'_N N\right)}{\bar{b}'_N N}\right)^{\frac{1}{d}}\right) \\
\leq\ & \mathbb{E}\left[P\left(h_{[0,1]^d}\left(\mathcal{X}'_N\right) > c_{h,1}\left(\frac{\log\left(\bar{b}'_N N\right)}{\bar{b}'_N N}\right)^{\frac{1}{d}} \mid N^{U'}, E\right) \mid E\right] + P\left(E^c\right) \\
\leq\ & \mathbb{E}\left[P\left(h_{[0,1]^d}\left(\mathcal{X}'_N\right) > c_{h,1}2^{-\frac{1}{d}}\left(\frac{\log\left(N^{U'}\right)}{N^{U'}}\right)^{\frac{1}{d}} \mid N^{U'}, E\right) \mid E\right] + o\left(\frac{1}{\bar{b}'_N N}\right) \qquad (2.24)
\end{aligned}$$

where (2.24) is due to Lemma 2.18 as well as (2.23).

When N satisfies

$$N \geq \min\left\{n\ :\ \bar{b}'_{n'} n' > e\ \forall n' \geq n\right\},$$

we have $N^{U'} \geq \frac{1}{2}\bar{b}'_N N + 1 > 1$ on the event E. Consequently, letting $c_{h,1} = 2^{\frac{1}{d}} c_h$, where

c_h is the constant obtained from Lemma 2.17, we obtain

$$\begin{aligned} & P\left(h_{[0,1]^d}\left(\mathcal{X}'_N\right) > c_{h,1}\left(\frac{\log\left(\bar{b}'_N N\right)}{\bar{b}'_N N}\right)^{\frac{1}{d}}\right) \\ \leq\ & \mathbb{E}\left(c_t \frac{1}{N^{U'}} \mid E\right) + o\left(\frac{1}{\bar{b}'_N N}\right) \qquad (2.25) \\ \leq\ & c_t \frac{1}{\frac{1}{2}\bar{b}'_N N + 1} + o\left(\frac{1}{\bar{b}'_N N}\right) \\ \leq\ & c_{t,1} \frac{1}{\bar{b}'_N N}, \end{aligned}$$

where (2.25) applies Lemma 2.17 to (2.24), and $c_{t,1}$ is suitably chosen to dominate the second term in (2.25). □

The concentration inequality can now be generalized to the domain $\mathcal{X}$. Since we assume that $\mathcal{X}$ is shape-regular, there exists a continuously differentiable bijection $L^{\mathcal{X}} : [0,1]^d \to \mathcal{X}$ whose Jacobian has nonzero determinant everywhere on its domain. The properties of this function are used in the proof.

Lemma 2.20. *There exist constants $c_{h,2}, c_{t,2} > 0$ such that, for all large enough N,*

$$P\left(h_{\mathcal{X}}\left(\mathcal{X}_N\right) > c_{h,2}\left(\frac{\log\left(bN\right)}{bN}\right)^{\frac{1}{d}}\right) \leq c_{t,2}\frac{1}{bN}. \qquad (2.26)$$

Proof: As in the proof of Lemma 2.5, we observe that the mapping $L_{\mathcal{X}}$ is Lipschitz. The Lipschitz condition implies

$$h_{\mathcal{X}}\left(\mathcal{X}_N\right) \leq c_{L_{\mathcal{X}}} h_{[0,1]^d}\left(L_{\mathcal{X}}^{-1}\left(\mathcal{X}_N\right)\right).$$

For each n, define a density

$$g'_n(y) = g_n(L_{\mathcal{X}}(y)) \left| \det\left(\frac{dL_{\mathcal{X}}(y)}{dy} \right) \right|.$$

Then, letting $b'_n = \inf_{y \in [0,1]^d} g'_n(y)$, we have $b'_n \geq b c_J$, where $c_J = \inf_{y \in [0,1]^d} \left| \det\left(\frac{dL_{\mathcal{X}}(y)}{dy} \right) \right|$. We also have $\bar{b}'_N \geq b c_J$, where $\bar{b}'_N = \frac{1}{N} \sum_{n=1}^{N} b'_n$.

For any arbitrary positive value of $c_{h,2}$, the inequalities

$$\begin{aligned} & P\left(h_{\mathcal{X}}(\mathcal{X}_N) > c_{h,2} \left(\frac{\log(bN)}{bN} \right)^{\frac{1}{d}} \right) \\ \leq \; & P\left(h_{[0,1]^d}\left(L_{\mathcal{X}}^{-1}(\mathcal{X}_N) \right) > \frac{c_{h,2}}{c_{L_{\mathcal{X}}}} \left(\frac{\log(bN)}{bN} \right)^{\frac{1}{d}} \right) \\ \leq \; & P\left(h_{[0,1]^d}\left(L_{\mathcal{X}}^{-1}(\mathcal{X}_N) \right) > \frac{c_{h,2}}{c_{L_{\mathcal{X}}}} c'_J \left(\frac{\log(\bar{b}'_N N)}{\bar{b}'_N N} \right)^{\frac{1}{d}} \right), \end{aligned}$$

where c'_J is some constant, hold for all N satisfying

$$N \geq \min\left\{ n \,:\, bn' > \max\left\{ e, \frac{e}{c_J} \right\} \; \forall n' \geq n \right\}. \tag{2.27}$$

Now choose $c_{h,2} = c_{h,1} \frac{c_{L_{\mathcal{X}}}}{c'_J}$, where $c_{h,1}$ is the constant obtained from Lemma 2.19. It follows that, for all N satisfying (2.27), we have

$$P\left(h_{\mathcal{X}}(\mathcal{X}_N) > c_{h,2} \left(\frac{\log(bN)}{bN} \right)^{\frac{1}{d}} \right) \leq c_{t,1} \frac{1}{\bar{b}'_N N} \leq \frac{c_{t,1}}{c_J} \frac{1}{bN},$$

where $c_{t,1}$ is the constant obtained from Lemma 2.19. Letting $c_{t,2} = \frac{c_{t,1}}{c_J}$ yields the desired result. □

Note that, if (2.26) holds for a certain value of $c_{h,2}$, it also holds for any larger value. The same is true of the other concentration inequalities on the mesh norm that we derived throughout this section.

2.7.3 Concentration inequality for estimation error

Finally, we connect the previously obtained results for the mesh norm to the estimation error $\|\hat{x}_N^* - x^*\|$. We first take care of the situation where $\hat{x}_N^*$ converging to x^* too slowly, by deriving a bound on the probability of this event. This is done by building on Lemma 2.10.

Then, conditional on the event $\{\hat{x}_N^* \in B(x^*, \rho_0)\}$, we apply (2.15), which implies the existence of a constant c_b' such that, for all N,

$$\|\hat{x}_N^* - x^*\| \leq c_b' h_{B(x^*,\rho_0)}^{\frac{k}{2}}(\mathcal{X}_N). \tag{2.28}$$

At this point, Lemma 2.20 provides a probabilistic bound for the right-hand side of (2.28). By combining this bound with a finite-time analysis of the probability that x_N^* is in a suitable neighborhood of x^*, we will obtain the final concentration inequality.

Lemma 2.21. *There exist constants $c_b, c_{t,3} > 0$ such that, for all large enough N,*

$$P\left(\hat{x}_N^* \in \mathcal{X} \setminus B\left(x^*, \min\left\{\rho_0, c_b\left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}}\right\}\right)\right) \leq c_{t,3}\frac{1}{bN}.$$

Proof: We derive

$$\begin{aligned}
& P\left(\hat{x}_N^* \in \mathcal{X} \setminus B\left(x^*, \min\left\{\rho_0, c_b\left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}}\right\}\right)\right) \\
= \; & P\left(\left\{\|\hat{x}_N^* - x^*\| > c_b\left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}}\right\} \cup \{\hat{x}_N^* \in \mathcal{X} \setminus B(x^*, \rho_0)\}\right) \\
\leq \; & P\left(\|\hat{x}_N^* - x^*\| > c_b\left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}} \mid \hat{x}_N^* \in B(x^*, \rho_0)\right) + P\left(\hat{x}_N^* \in \mathcal{X} \setminus B(x^*, \rho_0)\right) \\
\leq \; & P\left(\|\hat{x}_N^* - x^*\| > c_b\left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}} \mid \hat{x}_N^* \in B(x^*, \rho_0)\right) + o\left(\frac{1}{bN}\right),
\end{aligned}$$

where the last line follows from Lemma 2.10. We then derive

$$\begin{aligned}
& P\left(\|\hat{x}_N^* - x^*\| > c_b\left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}} \mid \hat{x}_N^* \in B(x^*, \rho_0)\right) \\
\leq \; & P\left(c_b' h_{\mathcal{X}}^{\frac{k}{2}}(\mathcal{X}_N) > c_b\left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}} \mid \hat{x}_N^* \in B(x^*, \rho_0)\right) \qquad (2.29)
\end{aligned}$$

where c_b' in (2.29) is the same value as in (2.28). Now, if we choose $c_b = c_b' c_{h,2}^{\frac{k}{2}}$, where $c_{h,2}$ is the same as in Lemma 2.20, we obtain

$$P\left(\|\hat{x}_N^* - x^*\| > c_b\left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}} \mid \hat{x}_N^* \in B(x^*, \rho_0)\right) \leq c_{t,2}\frac{1}{bN},$$

where $c_{t,2}$ is the same as in Lemma 2.20. The desired result follows. □

The next step of our analysis is to show that, if r_n is set according to (2.5) with some sufficiently large c_r, then the local sampling regions $S\left(\hat{x}_{n-1}^*, r_n\right)$ for $n \leq N$ will cover

$$S_N = S\left(x^*, c_b\left(\frac{\log(bN)}{bN}\right)^{\frac{k}{2d}}\right), \qquad (2.30)$$

where c_b is the value obtained from Lemma 2.21, with sufficiently high frequency as $N \to \infty$. In other words, there will be sufficiently many iterations n in which a rectangle centered at $\hat{x}_n^*$ will cover a rectangle centered at x^*, with the sizes of both rectangles shrinking as n grows large.

Before we proceed, we slightly relax the constant $c_{h,2}$ in Lemma 2.20. Note that, from (2.9), we have

$$h_{\mathcal{X}}\left(\mathcal{X}_N\right) = O\left(\left(\frac{\log\left(bN\right)}{bN}\right)^{\frac{1}{d}}\right)$$

almost surely. From Lemma 2.3, we know that there exists a *nonrandom* constant $c_{h,3}$ such that

$$h_{\mathcal{X}}\left(\mathcal{X}_N\right) \leq c_{h,3}\left(\frac{\log\left(bN\right)}{bN}\right)^{\frac{1}{d}} \tag{2.31}$$

for all large enough N (the exact threshold value of N at which this happens may be random, however). Thus, we can let $c_{h,4} = \max\left\{c_{h,2}, c_{h,3}\right\}$ and replace $c_{h,2}$ in Lemma 2.20 by $c_{h,4}$ without changing the result. Similarly, if N is large enough for (2.31) to hold, then we will still have inequality (2.31) if we replace $c_{h,3}$ by $c_{h,4}$.

Now, consider the set $\mathcal{X}_n$ of the first n data points. We observe that $h_{\mathcal{X}}\left(\mathcal{X}_n\right)$ is decreasing in n. Thus, if $\hat{x}_N^* \in B\left(x^*, \rho_0\right)$ and $h_{\mathcal{X}}\left(\mathcal{X}_{n'}\right) \leq c_{h,4}\left(\frac{\log(bn')}{bn'}\right)^{\frac{1}{d}}$ for some n',

then

$$\begin{aligned} \|\hat{x}_N^* - x^*\| &\leq c_b' h_{\mathcal{X}}^{\frac{k}{2}}(\mathcal{X}_n) && (2.32) \\ &\leq c_b' h_{\mathcal{X}}^{\frac{k}{2}}(\mathcal{X}_{n'}) \\ &\leq c_b' c_{h,4}^{\frac{k}{2}} \left(\frac{\log(bn')}{bn'}\right)^{\frac{k}{2d}} \\ &= c_b \left(\frac{\log(bn')}{bn'}\right)^{\frac{k}{2d}} && (2.33) \end{aligned}$$

for all $n' \leq n \leq N$. The value of c_b' in (2.32) is the same as in (2.28), and the value of c_b in (2.33) is obtained from Lemma 2.21 with $c_{h,4}$ replacing $c_{h,3}$ as discussed previously. Consequently, for suitably chosen c_r in (2.5), we have

$$S_N \subseteq S_{n'+1} \subseteq S(\hat{x}_{n'}^*, r_{n'+1})$$

where $S_N, S_{n'+1}$ are as in (2.30). It follows that the number of time stages $n \leq N$ in which S_N is covered by $S\left(\hat{x}_{n-1}^*, r_n\right)$ is equal to the number of times that

$$h_{\mathcal{X}}(\mathcal{X}_{n-1}) \leq c_{h,4} \left(\frac{\log(b(n-1))}{b(n-1)}\right)^{\frac{1}{d}}$$

is achieved.

Let n be some fixed integer large enough to satisfy the inequality (2.27), and define

$$z_{n'}^S = \begin{cases} 1 & h_{\mathcal{X}}(\mathcal{X}_{n'-1}) \leq c_{h,4} \left(\frac{\log(b(n'-1))}{b(n'-1)}\right)^{\frac{1}{d}}, \\ 0 & \text{otherwise,} \end{cases}$$

with $N^S = \sum_{n'=n}^{N} z_{n'}^S$. Thus, N^S counts the number of time stages $n \leq n' \leq N$ in which $S\left(\hat{x}_{n'-1}^*, r_{n'}\right)$ covers S_N. Recall that (2.31) holds for all large enough n', though the exact threshold after which this occurs may be random. Therefore, $z_{n'}^S = 1$ for all large enough n', whereupon

$$\liminf_{N' \to \infty} \frac{1}{N'} \sum_{n'=n}^{N'} z_{n'}^S = 1 \tag{2.34}$$

holds almost surely. We will discard a suitable set of measure zero from the outcome space so that (2.34) can be assumed to always hold and we do not have to keep conditioning on this event in our analysis.

With this, let $N \geq n$ be a fixed value satisfying (2.27). Then,

$$\begin{aligned} P\left(N^S = 0\right) &= P\left(h_{\mathcal{X}}\left(\mathcal{X}_{n'-1}\right) > c_{h,4}\left(\frac{\log\left(b\left(n'-1\right)\right)}{b\left(n'-1\right)}\right)^{\frac{1}{d}} \ \forall n \leq n' \leq N\right) \\ &\leq P\left(h_{\mathcal{X}}\left(\mathcal{X}_{N-1}\right) > c_{h,4}\left(\frac{\log\left(b\left(N-1\right)\right)}{b\left(N-1\right)}\right)^{\frac{1}{d}}\right) \\ &\leq c_{t,2}\frac{1}{b\left(N-1\right)} \end{aligned} \tag{2.35}$$

where the last line follows by Lemma 2.20.

Analogous to Section 2.6.2, we can view the distribution of design points under the ε-greedy policy as a mixture of uniform distributions. In those time stages n' where $z_{n'}^S = 1$ (i.e., where S_N is covered by the local search region), a portion of the design points can be viewed as originating from a uniform density on S_N. More precisely, we can let $z'_{n'}$, for $n \leq n' \leq N$, be independent Bernoulli random variables with success probabilities $t_{n'}\left|S_N\right|$. Then, $N^{S,U} = \sum_{n'=n}^{N} z_{n'}^S z'_{n'}$ is the cardinality of the subset $\mathcal{X}_N^{S,U}$ of the data that was sampled from a uniform distribution defined on S_N. Note that $z_{n'}^S$ and

$z'_{n'}$ are independent. We introduce the notation n_k, for $k = 1, 2, ..., N^S$, to represent those time stages n' for which $z^S_{n'} = 1$, that is,

$$n_k = \min \left\{ n' > n_{k-1} \, : \, z^S_{n'} = 1 \right\},$$

and use the notation $\mathcal{N} = \{n_k\}_{k=1}^{N^S}$ to denote the entire sequence of such time stages (all of which are random variables).

Recall from Lemma 2.17 that, if $\mathcal{X}_0^U$ is a set of i.i.d. samples from the uniform density on $[0,1]^d$, then

$$P\left(h_{[0,1]^d}\left(\mathcal{X}_0^U\right) > c_h \left(\frac{\log\left(\left|\mathcal{X}_0^U\right|\right)}{\left|\mathcal{X}_0^U\right|} \right)^{\frac{1}{d}} \right) \leq c_t \frac{1}{\left|\mathcal{X}_0^U\right|} \tag{2.36}$$

for some c_h, c_t. As we have observed previously in Section 2.6.2, when any set $\mathcal{D}$ is scaled isotropically and the sampling density is also appropriately scaled, the mesh norm will be changed by the same scaling factor. For this reason, we can apply inequality (2.36) to the mesh norm on S_N by treating it as a scaled mesh norm on $[0,1]^d$. That is, if we suppose that the value of $N^{S,U}$ is given, and let

$$E_N^x = \{\hat{x}_N^* \in B\left(x^*, \rho_0\right) \cap S_N\}$$

for notational convenience, we then have

$$P\left(h_{S_N}\left(\mathcal{X}_N^{S,U}\right) > c_h c_b \left(\frac{\log\left(bN\right)}{bN} \right)^{\frac{k}{2d}} \left(\frac{\log N^{S,U}}{N^{S,U}} \right)^{\frac{1}{d}} \mid N^{S,U},\ E_N^x,\ N^S \geq 1 \right) \leq \frac{c_t}{N^{S,U}}. \tag{2.37}$$

Recall from (2.19) that $t_n = b + \frac{1-b|\mathcal{X}|}{r_n^d} = O\left(\left(\frac{\log(bn)}{bn}\right)^{-\frac{k}{2}}\right)$. Now, we define

$$\bar{t}^S = \frac{1}{N^S}\sum_{k=1}^{N^S} t_{n_k},$$

and repeat the proof of Lemma 2.18 to obtain the inequality

$$\left(\frac{\log N^{S,U}}{N^{S,U}}\right)^{\frac{1}{d}} \leq 2^{\frac{1}{d}}\left(\frac{\log\left(\bar{t}^S \cdot |S_N| \cdot N^S\right)}{\bar{t}^S \cdot |S_N| \cdot N^S}\right)^{\frac{1}{d}} \tag{2.38}$$

under the condition that $N^{S,U} > \frac{1}{2}\bar{t}^S \cdot |S_N| \cdot N^S$. This allows us to improve the lower bound on $h_{S_N}\left(\mathcal{X}_N^{S,U}\right)$ in the event whose probability is computed in (2.37) to match the asymptotic a.s. rate of h_{S_N}. The resulting inequality can then be directly connected to the estimation error.

First, we give a technical lemma characterizing the probability that the condition required for (2.38) is *not* satisfied.

Lemma 2.22.

$$P\left(N^{S,U} < \frac{1}{2}\bar{t}^S\,|S_N|\,N^S + 1 \,|\, E_N^x, N^S \geq 1\right) = O\left(\frac{1}{(t_N\,|S_N|)^2\,N}\right).$$

Proof: Conditional on $\mathcal{N}$, $N^{S,U}$ is a sum of independent Bernoulli random variables.

Applying Hoeffding's concentration inequality, we derive

$$
\begin{aligned}
& P\left(N^{S,U} < \frac{1}{2}\bar{t}^S\,|S_N|\,N^S + 1 \,|\, E_N^x, N^S \geq 1\right) \\
\leq\ & P\left(\left|N^{S,U} - \bar{t}^S \cdot |S_N| \cdot N^S\right| > \frac{1}{2}\bar{t}^S\,|S_N|\,N^S - 1 \,|\, E_N^x, N^S \geq 1\right) \\
=\ & \mathbb{E}\left[P\left(\left|N^{S,U} - \bar{t}^S \cdot |S_N| \cdot N^S\right| > \frac{1}{2}\bar{t}^S\,|S_N|\,N^S - 1 \,|\, \mathcal{N}, E_N^x, N^S \geq 1\right)\right. \\
& \left. \,|\, E_N^x, N^S \geq 1\right] \\
\leq\ & 2\mathbb{E}\left[e^{-\frac{2}{N^S}\left(\frac{1}{2}\bar{t}^S|S_N|N^S - 1\right)^2} \,|\, E_N^x, N^S \geq 1\right] \\
\lesssim\ & \mathbb{E}\left[\frac{1}{\left(\bar{t}^S\,|S_N|\right)^2 N^S} \,|\, E_N^x, N^S \geq 1\right] \\
=\ & \mathbb{E}\left[\frac{1}{\left(\bar{t}^S\right)^2 N^S} \,|\, E_N^x, N^S \geq 1\right] \frac{1}{|S_N|^2}.
\end{aligned}
\tag{2.39}
$$

Provided that $N^S \geq 1$ and $\liminf_{N\to\infty} \frac{N^S}{N} = 1$, the inequality $N^S \leq n_{N^S}$ implies that $n_{N^S} \cong N$, whence, applying (2.19), we obtain

$$t_N \cong t_{n_{N^S}} \cong \bar{t}^S,$$

whence $\limsup_{N\to\infty} \frac{t_N}{\bar{t}^S} = 1$ and $\limsup_{N\to\infty} \frac{N}{N^S} = 1$. Therefore,

$$\limsup_{N\to\infty} \left(\frac{t_N}{\bar{t}^S}\right)^2 \frac{N}{N^S} \leq \limsup_{N\to\infty} \left(\frac{t_N}{\bar{t}^S}\right)^2 \cdot \limsup_{N\to\infty} \frac{N}{N^S} = 1.$$

Since $\bar{t}^S$ and N^S are all strictly positive random variables when $N^S \geq 1$, we can apply

Fatou's lemma to obtain

$$\begin{aligned}\limsup_{N\to\infty}\mathbb{E}\left[\frac{t_N^2 N}{\left(\bar{t}^S\right)^2 N^S}\,|\,E_N^x, N^S\geq 1\right] &\leq \mathbb{E}\left[\limsup_{N\to\infty}\frac{t_N^2 N}{\left(\bar{t}^S\right)^2 N^S}\,|\,E_N^x, N^S\geq 1\right]\\ &\leq 1,\end{aligned}$$

which yields a further bound on (2.39) due to the relation

$$\mathbb{E}\left[\frac{1}{\left(\bar{t}^S\right)^2 N^S}\,|\,E_N^x, N^S\geq 1\right]\lesssim\frac{1}{t_N^2 N},$$

for any N satisfying (2.27). This completes the proof. □

Now, letting $c_{h,5}=2^{\frac{1}{d}}c_h$, where c_h is the value obtained from (2.37), and also letting $E_N^S=\left\{N^S\geq 1\right\}$ for notational convenience, we can derive

$$\begin{aligned}&P\left(h_{S_N}\left(\mathcal{X}_N\right)>c_{h,5}\left(\frac{\log\left(\bar{t}^S\cdot|S_N|\cdot N^S\right)}{\bar{t}^S N^S}\right)^{\frac{1}{d}}\right.\\&\qquad\left.|\,\mathcal{N}, E_N^x, E_N^S, N^{S,U}>\frac{1}{2}\bar{t}^S\cdot|S_N|\cdot N^S\right)\\=\;&P\left(h_{S_N}\left(\mathcal{X}_N\right)>c_{h,5}c_b\left(\frac{\log\left(bN\right)}{bN}\right)^{\frac{k}{2d}}\left(\frac{\log\left(\bar{t}^S\cdot|S_N|\cdot N^S\right)}{\bar{t}^S\cdot|S_N|\cdot N^S}\right)^{\frac{1}{d}}\right.\\&\qquad\left.|\,\mathcal{N}, E_N^x, E_N^S, N^{S,U}>\frac{1}{2}\bar{t}^S\cdot|S_N|\cdot N^S\right)\\\leq\;&P\left(h_{S_N}\left(\mathcal{X}_N\right)>2^{-\frac{1}{d}}c_{h,5}c_b\left(\frac{\log\left(bN\right)}{bN}\right)^{\frac{k}{2d}}\left(\frac{\log N^{S,U}}{N^{S,U}}\right)^{\frac{1}{d}}\right.\\&\qquad\left.|\,\mathcal{N}, E_N^x, E_N^S, N^{S,U}>\frac{1}{2}\bar{t}^S\cdot|S_N|\cdot N^S\right)\qquad(2.40)\\=\;&P\left(h_{S_N}\left(\mathcal{X}_N\right)>c_h c_b\left(\frac{\log\left(bN\right)}{bN}\right)^{\frac{k}{2d}}\left(\frac{\log N^{S,U}}{N^{S,U}}\right)^{\frac{1}{d}}\right.\\&\qquad\left.|\,\mathcal{N}, E_N^x, E_N^S, N^{S,U}>\frac{1}{2}\bar{t}^S\cdot|S_N|\cdot N^S\right),\qquad(2.41)\end{aligned}$$

where (2.40) is obtained by applying (2.38), while (2.41) follows from the definition of $c_{h,5}$.

We then derive the bound

$$
\begin{aligned}
& P\left(h_{S_N}\left(\mathcal{X}_N\right) > c_{h,5}\left(\frac{\log\left(\bar{t}^S\cdot|S_N|\cdot N^S\right)}{\bar{t}^S N^S}\right)^{\frac{1}{d}} \mid E_N^x, E_N^S\right) \\
= \;& \mathbb{E}\left[P\left(h_{S_N}\left(\mathcal{X}_N\right) > c_{h,5}\left(\frac{\log\left(\bar{t}^S\cdot|S_N|\cdot N^S\right)}{\bar{t}^S N^S}\right)^{\frac{1}{d}} \mid \mathcal{N}, E_N^x, E_N^S\right) \mid E_N^x, E_N^S\right] \\
\leq \;& \mathbb{E}\left[P\left(h_{S_N}\left(\mathcal{X}_N\right) > c_{h,5}\left(\frac{\log\left(\bar{t}^S\cdot|S_N|\cdot N^S\right)}{\bar{t}^S N^S}\right)^{\frac{1}{d}}\right.\right. \\
& \left.\mid \mathcal{N}, E_N^x, E_N^S, N^{S,U} > \frac{1}{2}\bar{t}^S\cdot|S_N|\cdot N^S\right) \\
& \left.+P\left(N^{S,U} < \frac{1}{2}\bar{t}^S\cdot|S_N|\cdot N^S + 1 \mid \mathcal{N}, E_N^x, E_N^S\right) \mid E_N^x, E_N^S\right] \qquad (2.42) \\
\leq \;& \mathbb{E}\left[P\left(h_{S_N}\left(\mathcal{X}_N\right) > c_h c_b\left(\frac{\log\left(bN\right)}{bN}\right)^{\frac{k}{2d}}\left(\frac{\log N^{S,U}}{N^{S,U}}\right)^{\frac{1}{d}}\right.\right. \\
& \left.\left.\mid \mathcal{N}, E_N^x, E_N^S, N^{S,U} > \frac{1}{2}\bar{t}^S\cdot|S_N|\cdot N^S\right) \mid E_N^x, E_N^S\right] \\
& +O\left(\frac{1}{\left(t_N|S_N|\right)^2 N}\right) \qquad (2.43) \\
\leq \;& c_t\mathbb{E}\left[\mathbb{E}\left[\frac{1}{N^{S,U}} \mid E_N^x, E_N^S, N^{S,U} > \frac{1}{2}\bar{t}^S\cdot|S_N|\cdot N^S\right] \mid E_N^x, E_N^S\right] \\
& +O\left(\frac{1}{\left(t_N|S_N|\right)^2 N}\right) \qquad (2.44) \\
\leq \;& \frac{2c_t}{|S_N|}\mathbb{E}\left[\frac{1}{\bar{t}^S N^S} \mid E_N^x, E_N^S\right] + O\left(\frac{1}{\left(t_N|S_N|\right)^2 N}\right).
\end{aligned}
$$

In this derivation, (2.43) is obtained by applying (2.41) to the first term of (2.42), and Lemma 2.22 to the second term. Then, (2.44) is due to (2.37).

By repeating the proof of Lemma 2.22, we can derive

$$\mathbb{E}\left[\frac{1}{\bar{t}^S N^S} \mid E_N^x, E_N^S\right] \lesssim \frac{1}{t_N N}$$

for any N satisfying (2.27). Because $t_N |S_N| = O(1)$, this yields

$$\begin{aligned} & P\left(h_{S_N}(\mathcal{X}_N) > c_{h,5}\left(\frac{\log\left(\bar{t}^S \cdot |S_N| \cdot N^S\right)}{\bar{t}^S N^S}\right)^{\frac{1}{d}} \mid E_N^x, E_N^S\right) \\ = \; & O\left(\frac{1}{t_N \cdot |S_N| \cdot N}\right) + O\left(\frac{1}{(t_N \cdot |S_N|)^2 \cdot N}\right) \\ = \; & O\left(\frac{1}{N}\right). \end{aligned}$$

Similarly to (2.18), when $\hat{x}_N^* \in B(x^*, \rho_0) \cap S_N$, we can bound the error of $\hat{x}_N^*$ by the local mesh norm on S_N as

$$\|\hat{x}_N^* - x^*\| = O\left(h_{S_N}^{\frac{k}{2}}(\mathcal{X}_N)\right).$$

It follows from the preceding that there exists a constant $c_{b,1}$ such that

$$P\left(\|\hat{x}_N^* - x^*\| > c_{b,1}\left(\frac{\log\left(\bar{t}^S \cdot |S_N| \cdot N^S\right)}{\bar{t}^S N^S}\right)^{\frac{k}{2d}} \mid E_N^x, E_N^S\right) = O\left(\frac{1}{N}\right).$$

Using the fact that $t_N \cong \bar{t}^S$ and $N^S \cong N$, there exists a constant $c_{b,2}$ such that

$$\begin{aligned}
& P\left(\|\hat{x}_N^* - x^*\| > c_{b,2}\left(\frac{\log\left(t_N \cdot |S_N| \cdot N\right)}{t_N N}\right)^{\frac{k}{2d}}\right) \\
\leq\ & P\left(\|\hat{x}_N^* - x^*\| > c_{b,2}\left(\frac{\log\left(t_N \cdot |S_N| \cdot N\right)}{t_N N}\right)^{\frac{k}{2d}} \mid E_N^x, E_N^S\right) \\
& + P\left(1 - E_N^x\right) + P\left(1 - E_N^S\right) \\
\leq\ & P\left(\|\hat{x}_N^* - x^*\| > c_{b,1}\left(\frac{\log\left(\bar{t}^S \cdot |S_N| \cdot N^S\right)}{\bar{t}^S N^S}\right)^{\frac{k}{2d}} \mid E_N^x, E_N^S\right) + O\left(\frac{1}{bN}\right) \\
=\ & O\left(\frac{1}{N}\right),
\end{aligned}$$

shown by applying Lemma 2.21 together with (2.35). By combining this bound with the preceding analysis on the rate of t_N, we obtain constants c, c' for which

$$P\left(\|\hat{x}_N^* - x^*\| > c\left(\frac{\log N}{N}\right)^{\frac{k}{2d}}\left(\frac{\log\left(bN\right)}{bN}\right)^{\frac{k^2}{2d}}\right) \leq \frac{c'}{N},$$

for sufficiently large finite N.

2.8 Appendix: proofs

In this section, we provide full proofs for all results that were stated in the main text.

2.8.1 Proof of Lemma 2.8

First, observe that $\hat{x}_N^* \in B(x^*, \rho_0)$ if

$$\sup_{x \in \mathcal{X} \setminus B(x^*, \rho_0)} \left| \hat{f}_N(x) - f(x) \right| < f(\tilde{x}) - \min_{1 \le n \le N} f(x_n), \tag{2.45}$$

where $\tilde{x} = \arg\min_{x \in \mathrm{cl}(\mathcal{X} \setminus B(x^*, \rho_0))} f(x)$. This is because for a point outside $B(x^*, \rho_0)$ to be the global minimizer, the interpolant value at this point has to dive deep at least below the lowest response, which means a big interpolation error.

The inequality (2.45) is implied if

$$\sup_{x \in B(x_\rho, \rho)} \left| \hat{f}_N(x) - f(x) \right| < f(\tilde{x}) - \min_{1 \le n \le N} f(x_n) \tag{2.46}$$

holds for any x_ρ, ρ satisfying $B(x_\rho, \rho) \subseteq \mathrm{cl}(\mathcal{X} \setminus B(x^*, \rho_0))$. Note that the sufficient event (2.46) implicitly excludes the situation where no design points are in $B(x^*, \rho_0)$.

Next, by Lemma 2.1, for all large enough N and all x_ρ, ρ satisfying $B(x_\rho, \rho) \subseteq \mathrm{cl}(\mathcal{X} \setminus B(x^*, \rho_0))$, we have

$$\sup_{x \in B(x_\rho, \rho)} \left| \hat{f}_N(x) - f(x) \right| \le c_{f,\phi} C h^k_{B(x_\rho, \rho)}(\mathcal{X}_N) \le c_{f,\phi} C h^k_{\mathcal{X}}(\mathcal{X}_N),$$

where $h_{\mathcal{X}}(\mathcal{X}_N)$ does not depend on x_ρ, ρ. Thus, the sufficient condition (2.46) is attained

if

$$\begin{aligned} c_{f,\phi} C h_{\mathcal{X}}^{k}\left(\mathcal{X}_N\right) &< f\left(\tilde{x}\right) - \min_{1\leq n\leq N} f\left(x_n\right) \\ &= f\left(\tilde{x}\right) - f\left(x^*\right) - \left(\min_{1\leq n\leq N} f\left(x_n\right) - f\left(x^*\right)\right), \end{aligned}$$

or, equivalently,

$$c_{f,\phi} C h_{\mathcal{X}}^{k}\left(\mathcal{X}_N\right) + \left(\min_{1\leq n\leq N} f\left(x_n\right) - f\left(x^*\right)\right) < f\left(\tilde{x}\right) - f\left(x^*\right). \tag{2.47}$$

Now, define

$$\begin{aligned} c_w &= \left(\frac{1}{2c_{f,\phi}C}\left(f\left(\tilde{x}\right) - f\left(x^*\right)\right)\right)^{\frac{1}{k}}, \\ \mathcal{D} &= \left\{x \in \mathcal{X} \,:\, f\left(x\right) < \frac{1}{2}\left(f\left(\tilde{x}\right) + f\left(x^*\right)\right)\right\}. \end{aligned}$$

On the event $\left\{h_{\mathcal{X}}\left(\mathcal{X}_N\right) < c_w\right\} \cap \bigcup_{n=1}^{N}\left\{x_n \in \mathcal{D}\right\}$, the inequality (2.47) is attained, which, in turn, implies $\hat{x}_*^N \in B\left(x^*, \rho_0\right)$ as desired.

Because $\mathcal{X}_N^U \subseteq \mathcal{X}$, we can easily see that $h_{\mathcal{X}}\left(\mathcal{X}_N\right) \leq h_{\mathcal{X}}\left(\mathcal{X}_N^U\right)$, and hence it is sufficient to restrict ourselves to the event $\left\{h_{\mathcal{X}}\left(\mathcal{X}_N^U\right) < c_w\right\} \cap \bigcup_{n=1}^{N}\left\{x_n \in X_N \cap \mathcal{D}\right\}$. So, we conclude that

$$\begin{aligned} P\left(\hat{x}_N^* \in B\left(x^*, \rho_0\right)\right) &\geq P\left(h_{\mathcal{X}}\left(X_N^U\right) < c_w,\ \exists x_n \in \mathcal{X}_N \cap \mathcal{D}\right) \\ &\geq P\left(h_{\mathcal{X}}\left(\mathcal{X}_N^U\right) < c_w\right) + P\left(\bigcup_{n=1}^{N}\left\{x_n \in \mathcal{X}_N \cap \mathcal{D}\right\}\right) - 1, \end{aligned}$$

as required.

2.8.2 Proof of Lemma 2.9

By Lemma 2.5, we have

$$h_{\mathcal{X}}\left(\mathcal{X}_N^U\right) = O\left(\left(\frac{\log N^U}{N^U}\right)^{\frac{1}{d}}\right) \qquad \text{a.s.}$$

Therefore, with probability 1, there exists a finite but random value $c_{w,h}$ that

$$h_{\mathcal{X}}\left(\mathcal{X}_N^U\right) \leq c_{w,h}\left(\frac{\log N^U}{N^U}\right)^{\frac{1}{d}}, \qquad N^U > 1.$$

It follows that

$$\begin{aligned} c_{w,h}\left(N^U\right)^{-\frac{1-\varepsilon_{w,h}}{d}} < c_w \qquad &\Rightarrow \qquad c_{w,h}\left(\frac{\log N^U}{N^U}\right)^{\frac{1}{d}} < c_w \\ &\Rightarrow \qquad h_{\mathcal{X}}\left(\mathcal{X}_N^U\right) < c_w \end{aligned}$$

for any $\varepsilon_{w,h} \in (0,1)$ when $N^U > N_0$. Taking $\varepsilon_{w,h} = \frac{1}{2}$, we let $c'_w = \left(\frac{c_{w,h}}{c_w}\right)^{2d}$ which is also random and obtain $h_{\mathcal{X}}\left(\mathcal{X}_N^U\right) < c_w$ when $N^U > c'_w$.

Consequently,

$$\begin{aligned} P\left(h_{\mathcal{X}}\left(\mathcal{X}_N^U\right) < c_w\right) &\geq P\left(h_{\mathcal{X}}\left(\mathcal{X}_N^U\right) < c_w \,|\, N^U > N_0\right) P\left(N^U > N_0\right) \\ &\geq P\left(N^U > c'_w \,|\, N^U > N_0\right) P\left(N^U > N_0\right) \\ &= P\left(N^U > \max\left\{c'_w, N_0\right\}\right). \end{aligned} \tag{2.48}$$

To bound (2.48), we require a concentration inequality for N^U. Note that $N^U = \sum_{n=1}^{N} z_n$,

where z_n are i.i.d. $Bernoulli\left(b\left|\mathcal{X}\right|\right)$ random variables. The expected value of N^U increases to infinity with N.

We derive

$$\begin{aligned}
& P\left(N^U > \max\left\{c'_w, 1\right\}\right) \\
= \;& P\left(N^U - \mathbb{E}\left(N^U\right) > -\left(\mathbb{E}\left(N^U\right) - \max\left\{c'_w, 1\right\}\right)\right) \\
= \;& 1 - P\left(N^U - \mathbb{E}\left(N^U\right) < -\left(\mathbb{E}\left(N^U\right) - \max\left\{c'_w, 1\right\}\right)\right) \\
\geq \;& 1 - P\left(N^U - \mathbb{E}\left(N^U\right) < -\left(\mathbb{E}\left(N^U\right) - \max\left\{c'_w, 1\right\}\right) \left| \frac{1}{2}\mathbb{E}\left(N^U\right) > \max\left\{c'_w, 1\right\}\right.\right) \\
\geq \;& 1 - P\left(\left|N^U - \mathbb{E}\left(N^U\right)\right| > \mathbb{E}\left(N^U\right) - \max\left\{c'_w, 1\right\} \left| \frac{1}{2}\mathbb{E}\left(N^U\right) > \max\left\{c'_w, 1\right\}\right.\right).
\end{aligned}$$

Since, by Hoeffding's concentration inequality, we can obtain

$$\begin{aligned}
& P\left(\left|N^U - \mathbb{E}\left(N^U\right)\right| > \mathbb{E}\left(N^U\right) - \max\left\{c'_w, N_0\right\} \left| c'_w\right.\right) \\
\leq \;& 2\exp\left\{-\frac{2}{N}\left(\mathbb{E}\left(N^U\right) - \max\left\{c'_w, 1\right\}\right)^2\right\},
\end{aligned}$$

we get

$$\begin{aligned}
& P\left(\left|N^U - \mathbb{E}\left(N^U\right)\right| > \mathbb{E}\left(N^U\right) - \max\left\{c'_w, N_0\right\} \left| \frac{1}{2}\mathbb{E}\left(N^U\right) > \max\left\{c'_w, 1\right\}\right.\right) \\
\leq \;& 2\exp\left\{-\frac{2}{N}\left(\frac{b\left|\mathcal{X}\right|N}{2}\right)^2\right\} \\
\lesssim \;& e^{-b^2|\mathcal{X}|^2 N/2}.
\end{aligned}$$

The desired result follows.

2.8.3 Proof of Lemma 2.11

For any vector $v \neq 0$ in $\mathbb{R}^d$,

$$\begin{aligned}
\frac{v^\top A v}{\|v\|^2} &\leq \frac{\left(\sum_i \sum_j A_{ij}^2\right)^{\frac{1}{2}} \left(\sum_i \sum_j v_i^2 v_j^2\right)^{\frac{1}{2}}}{\|v\|^2} \\
&\leq \frac{(d^2 t_A^2)^{\frac{1}{2}} \left(\sum_j v_j^2\right)}{\|v\|^2} \\
&= d \cdot t_A.
\end{aligned}$$

The desired result follows by taking the supremum over all nonzero v.

2.8.4 Proof of Lemma 2.13

We prove this result by induction. When $m = 1$, it is obviously correct; when $m = 2$, we have $P\left(W\left(N, 2\right) \succeq \delta\right) = 1 - \frac{1}{2^{N-1}}$, also according to (2.20).

Suppose now that (2.20) holds for $W(N, m')$, $m' = 1, ..., m$. Then,

$$
\begin{aligned}
& P\left(W\left(N, m+1\right) \succeq \delta\right) \\
= \; & 1 - \sum_{k=1}^{m} P\left(\sum_{j} 1_{\{W_j > 0\}} = k\right) \\
= \; & 1 - \sum_{k=1}^{m} \binom{m+1}{k} \left(\frac{k}{m+1}\right)^{N} P\left(W\left(N, k\right) \succeq \delta\right) \\
= \; & 1 - \frac{1}{\left(m+1\right)^{N}} \left[\sum_{k=1}^{m} \binom{m+1}{k} k^{N} + \sum_{k=2}^{m} \binom{m+1}{k} \sum_{j=1}^{k-1} \binom{k}{j} j^{N} \left(-1\right)^{k-j}\right] \\
= \; & 1 - \frac{1}{\left(m+1\right)^{N}} \left[\sum_{k=1}^{m} \binom{m+1}{k} k^{N} + \sum_{j=1}^{m-1} j^{N} \sum_{k=j+1}^{m} \binom{m+1}{k} \binom{k}{j} \left(-1\right)^{k-j}\right] \\
= \; & 1 - \frac{1}{\left(m+1\right)^{N}} \left[\sum_{k=1}^{m} \binom{m+1}{k} k^{N} + \sum_{j=1}^{m-1} j^{N} \binom{m+1}{j} \sum_{k=j+1}^{m} \binom{m+1-j}{k-j} \left(-1\right)^{k-j}\right] \\
= \; & 1 - \frac{1}{\left(m+1\right)^{N}} \left[\sum_{k=1}^{m} \binom{m+1}{k} k^{N} + \sum_{j=1}^{m-1} j^{N} \binom{m+1}{j} \sum_{k=1}^{m-j} \binom{m-j+1}{k} \left(-1\right)^{k}\right] \\
= \; & 1 - \frac{1}{\left(m+1\right)^{N}} \left[\sum_{k=1}^{m} \binom{m+1}{k} k^{N} + \sum_{j=1}^{m-1} j^{N} \binom{m+1}{j} \left(\left(-1\right)^{m-j} - 1\right)\right] \\
= \; & 1 - \frac{1}{\left(m+1\right)^{N}} \left[\binom{m+1}{m} m^{N} + \sum_{j=1}^{m-1} \binom{m+1}{j} j^{N} \left(-1\right)^{m-j}\right] \\
= \; & 1 - \frac{1}{\left(m+1\right)^{N}} \sum_{j=1}^{m} \binom{m+1}{j} j^{N} \left(-1\right)^{m-j} \\
= \; & \frac{1}{\left(m+1\right)^{N}} \sum_{j=0}^{m+1} \binom{m+1}{j} j^{N} \left(-1\right)^{m-j+1},
\end{aligned}
$$

thus verifying (2.20) and completing the proof.

2.8.5 Proof of Lemma 2.15

Let y_0 be the solution of the equation $\frac{m}{n}y = 1 - e^{-y}$, and let $s = \frac{n}{m}$. Then, we have

$$\begin{aligned}
\frac{y_0 - s}{s} &= -e^{-y_0} \\
(y_0 - s)\, e^{y_0} &= -s \\
(y_0 - s)\, e^{y_0 - s} &= -se^{-s}.
\end{aligned}$$

This implies $y_0 = s - w$, where $w = -W_0\left(-se^{-s}\right)$. From Lemma 2.14 we have (2.21), and we can derive

$$\begin{aligned}
e^T m^{n-m} \tau\left(s_0\right) \binom{n}{m} &= e^{\gamma(y_0) - n + m + (n-m)\log(s-1)} m^{n-m} \sqrt{\frac{s-1}{s\left(x_0 - s + 1\right)}} \binom{n}{m} \\
&= e^{m-n} \left(\frac{s-1}{y_0}\right)^{n-m} \frac{m^n}{\left(m\left(s - y_0\right)\right)^m} \sqrt{\frac{s-1}{s\left(x_0 - s + 1\right)}} \binom{n}{m} \\
&= e^{m-n} \left(\frac{s-1}{s-w}\right)^{n-m} \frac{m^{n-m}}{w^m} \sqrt{\frac{s-1}{s\left(1-w\right)}} \binom{n}{m},
\end{aligned}$$

as required.

2.8.6 Proof of Lemma 2.16

Let T and τ be as in the statement of Lemma 2.14, and let w be as in the statement of Lemma 2.15. Also let $s = \frac{N}{m}$. First, we derive

$$\begin{aligned} P(M < m) &= 1 - \frac{m!}{m^N} \left\{ {N \atop m} \right\} \\ &= 1 - \frac{m!}{m^N} e^T m^{N-m} \tau \left(\frac{N}{m} - 1 \right) \binom{N}{m} \\ &\quad + \frac{m!}{m^N} e^T m^{N-m} \tau \left(\frac{N}{m} - 1 \right) \binom{N}{m} - \frac{m!}{m^N} \left\{ {n \atop m} \right\} \\ &\leq \left| 1 - \frac{m!}{m^N} e^{m-N} \left(\frac{s-1}{s-w} \right)^{N-m} \frac{m^{N-m}}{w^m} \sqrt{\frac{s-1}{s(1-w)}} \binom{N}{m} \right| \\ &\quad + \left| \frac{m!}{m^N} \left(e^T m^{N-m} \tau (s-1) \binom{N}{m} - \left\{ {N \atop m} \right\} \right) \right|. \end{aligned} \tag{2.49}$$

Consider the decreasing rate of the first term on the right-hand side of (2.49). Taking advantage of Stirling's formula

$$n! = \sqrt{2\pi n} \left(\frac{n}{e} \right)^n \left(1 + O\left(\frac{1}{n} \right) \right),$$

we derive

$$
\begin{aligned}
& \left| 1-\frac{m!}{m^N} e^{m-N} \left(\frac{s-1}{s-w}\right)^{N-m} \frac{m^{N-m}}{w^m} \sqrt{\frac{s-1}{s\,(1-w)}} \binom{N}{m} \right| \\
= & \left| 1-\frac{e^{m-N}}{m^N} \left(\frac{1}{s-w}\right)^{N-m} (N-m)^{N-m} \frac{1}{w^m} \sqrt{\frac{N-m}{N\,(1-w)}} \frac{N!}{(N-m)!} \right| \\
= & \left| 1-\frac{N!}{m^N} \left(\frac{1}{s-w}\right)^{N-m} \frac{1}{w^m} \sqrt{\frac{1}{N\,(1-w)}} \frac{\sqrt{N-m}\left(\frac{N-m}{e}\right)^{N-m}}{(N-m)!} \right| \\
= & \left| 1-\frac{N!}{m^N} \left(\frac{1}{s-w}\right)^{N-m} \frac{1}{w^m} \sqrt{\frac{1}{N\,(1-w)}} \frac{1}{\sqrt{2\pi}} \left(1+O\left(\frac{1}{N-m}\right)\right) \right| \\
= & \left| 1-\frac{N!}{m^N} \left(\frac{1}{s-se^{-s}-O\left(s^2e^{-2s}\right)}\right)^{N-m} \frac{1}{\left(se^{-s}+O\left(s^2e^{-2s}\right)\right)^m} \frac{1}{\sqrt{2\pi N}} \right. \\
& \left. \cdot \sqrt{\frac{1}{1-se^{-s}-O\left(s^2e^{-2s}\right)}} \left(1+O\left(\frac{1}{N-m}\right)\right) \right| \qquad (2.50) \\
= & \left| 1-\frac{N!}{m^N} \frac{1}{s^N e^{-sm}} \left(\frac{1}{1-e^{-s}-O\left(se^{-2s}\right)}\right)^{N-m} \frac{1}{\left(1+O\left(se^{-s}\right)\right)^m} \frac{1}{\sqrt{2\pi N}} \right. \\
& \left. \cdot \sqrt{\frac{1}{1-se^{-s}-O\left(s^2e^{-2s}\right)}} \left(1+O\left(\frac{1}{N-m}\right)\right) \right| \\
= & \left| 1-\frac{N!}{\left(\frac{N}{e}\right)^N \sqrt{2\pi N}} \left(1+O\left(\left(N-m\right)e^{-s}\right)\right) \left(1+O\left(Ne^{-s}\right)\right) \left(1+O\left(se^{-s}\right)\right) \right. \\
& \left. \cdot \left(1+O\left(\frac{1}{N-m}\right)\right) \right| \\
= & \left| 1-\left(1+O\left(\frac{1}{N}\right)\right) \left(1+O\left(\frac{1}{N-m}\right)\right) \right| \\
= & \; O\left(\frac{1}{N-m}\right),
\end{aligned}
$$

where (2.50) is obtained by first using the Lagrange inversion theorem to derive the expansion

$$W_0\left(y\right) = \sum_{k=1}^{\infty} \frac{\left(-k\right)^{k-1}}{k!} y^k = y - \frac{1}{2}y^2 + O\left(y^3\right)$$

for the upper branch W_0 of the Lambert W function when $y \to 0^-$, and then observing that

$$w = se^{-s} + O\left(s^2 e^{-2s}\right) \tag{2.51}$$

for large s.

Now consider the decreasing rate of the second term on the right-hand side of (2.49). Based on Temme [1993], one can show that

$$\left\{ {n \atop m} \right\} - e^T m^{n-m} \tau(s-1) \binom{n}{m} \cong -e^T m^{n-m} \binom{n}{m} \frac{\tau_1(s-1)}{m} \tag{2.52}$$

whereupon, with some tedious algebra, one finds $\tau_1(s-1) \cong \frac{1}{3\tau(s-1)s}$. Then,

$$\begin{aligned}
& \left| \frac{m!}{m^N} \left(e^T m^{N-m} \tau(s-1) \binom{N}{m} - \left\{ {N \atop m} \right\} \right) \right| \\
\cong\ & \frac{m!}{m^N} e^T m^{N-m} \binom{N}{m} \frac{\tau_1(s-1)}{m} && (2.53) \\
=\ & \frac{e^T}{m^{m+1}} \frac{N!}{(N-m)!} \tau_1(s-1) \\
=\ & \frac{1}{m^{m+1}} \frac{N!}{(N-m)!} \left(\frac{1}{s-w} \right)^{N-m} \frac{(N-m)^{N-m}}{m^{N-m}} \frac{e^{m-N}}{w^m} \tau_1(s-1) && (2.54) \\
\cong\ & \frac{1}{m^{N+1}} \frac{N!}{(N-m)!} \left(\frac{N-m}{e} \right)^{N-m} \left(\frac{1}{s-w} \right)^{N-m} \frac{1}{w^m} \frac{1}{3\tau(s-1)s} \\
\cong\ & \frac{N!}{m^{N+1}} \frac{1}{\sqrt{2\pi(N-m)}} \frac{1}{(s-se^{-s})^{N-m}} \frac{1}{s^m e^{-N}} \frac{1}{3s} \sqrt{\frac{N}{N-m}} && (2.55) \\
=\ & \frac{1}{3(N-m)} \frac{N!}{N^N e^{-N} \sqrt{2\pi N}} \frac{1}{(1-e^{-s})^{N-m}} \\
\cong\ & \frac{1}{3(N-m)},
\end{aligned}$$

where (2.53) follows by (2.52), equation (2.54) follows from $e^T = e^{m-N} \frac{1}{w^m} \left(\frac{s-1}{s-w} \right)^{N-m}$, and (2.55) follows from (2.51) and the fact that $\tau(s-1) = \sqrt{\frac{s-1}{s(1-w)}}$. Since we have

found that both terms on the right-hand side of (2.49) are $O\left(\frac{1}{N-m}\right)$, the desired result follows.

2.8.7 Proof of Lemma 2.18

For $0 < z < y$ and $y - z > 1$, we have the inequality

$$\frac{y}{\log y} \leq \frac{y-z}{\log(y-z)} + \frac{z}{\log(y-z)}.$$

Take $y = \bar{b}'_N N$ and $z = \frac{1}{2}\bar{b}'_N N - 1$, divide both sides by $\frac{y-z}{\log(y-z)}$ and we can conclude

$$\begin{aligned}
\frac{\bar{b}'_N N}{\log\left(\bar{b}'_N N\right)} \frac{\log N^{U'}}{N^{U'}} &\leq \frac{\bar{b}'_N N}{\log\left(\bar{b}'_N N\right)} \frac{\log\left(\frac{1}{2}\bar{b}'_N N + 1\right)}{\frac{1}{2}\bar{b}'_N N + 1} \\
&\leq 1 + \frac{\frac{1}{2}\bar{b}'_N N - 1}{\frac{1}{2}\bar{b}'_N N + 1} \\
&\leq 2,
\end{aligned}$$

which leads to the desired result.

Chapter 3: Moderate deviations inequalities for Gaussian process regression

3.1 Introduction

Given a compact domain $D \subseteq \mathbb{R}^d$, let $\{\mathcal{E}(x)\}_{x \in D}$ be a centered Gaussian process (Gaussian random field) on a probability space $(\Omega, \mathcal{F}, P)$. Define

$$f(x) = m(x) + \mathcal{E}(x), \qquad x \in D, \tag{3.1}$$

where $m : D \to \mathbb{R}$ is a pre-specified unknown "mean function." Suppose that we are given the values $f(x_1), ..., f(x_n)$ of f at the *design points* $x_1, ..., x_n \in D$. Then, we can construct an estimator $\hat{f}_n$ of f using Gaussian process regression [Rasmussen and Williams, 2006]. This is a Bayesian method when we take the unconditinoal Gaussian process as a prior: for each x, $\hat{f}_n(x)$ is the conditional mean of the random variable $f(x)$ given $f(x_1), ..., f(x_n)$. The covariance function of the Gaussian process, assumed known, is used to infer the value of f at unobserved x from information collected about the design points.

Gaussian process regression is widely used to interpolate and predict the values of black-box functions in simulation calibration [Scott et al., 2010] and optimization [Jones

et al., 1998, Ankenman et al., 2010], biomedical applications [Lee et al., 2014], risk assessment of civil infrastructure [Sheibani and Ou, 2021], tuning of machine learning models [Snoek et al., 2012], and many other problems from diverse branches of science. In all such applications, f models the output of a complex system (physical or virtual), with x being the input. There is no closed form for f, but it is possible to observe $f(x)$ at individual x values, e.g., by running expensive lab, field, or computer experiments with those particular inputs. The goal is to obtain accurate estimates at unobserved values using as few experiments as possible. Often, the function f represents a performance metric, such as the predictive power of a machine learning model with a given set of parameters, and the goal then becomes to optimize $f(x)$ over $x \in D$.

The analysis of this chapter is motivated by concerns that arise in design of experiments, though we do not explicitly model any design problem. Our main contribution is a theoretical framework for studying the large deviations behavior of random vectors of the form $\left(\hat{f}_n(x), \hat{f}_n(x^*), f(x), f(x^*)\right)$ for two fixed but arbitrary points $x, x^* \in D$. This framework can be applied to prove new convergence rates for different types of "error probabilities" related to GP regression. We demonstrate the usefulness of the theory with two specific applications, though others may be possible. The first application deals with probabilities of the form

$$\pi_n(x, x^*) = P\left(\hat{f}_n(x) \leq \hat{f}_n(x^*) - \delta \mid f(x) \geq f(x^*)\right), \tag{3.2}$$

where $\delta > 0$ is a small threshold. In words, it is given to us that x^* has a smaller function value than x, but interpolation error may cause us to falsely reverse this ordering (the

threshold δ makes (3.2) well-defined). When f is an objective function, this is the probability of reporting x as being "better" than x^* when in reality the opposite is the case. For this type of error probability, we leverage our theory to prove a new moderate deviations inequality

$$P\left(\hat{f}_n(x) \leq \hat{f}_n(x) - \delta \mid f(x) \geq f(x^*)\right) \lesssim \exp\left(-\delta^2 C h_n^{-\frac{1}{2}s}\right) \tag{3.3}$$

where $C, s > 0$ are constants depending on the specification of the Gaussian process, and

$$h_n = \max_{y \in D} \min_{m=1,\ldots,n} \|y - x_m\|_2$$

is the *mesh norm* measuring the density of the design points. In a special case where the design points are uniformly distributed on D, it has been shown [Janson, 1987] that h_n is of order $\left(\frac{\log n}{n}\right)^{\frac{1}{d}}$, which justifies the interpretation of (3.3) as a moderate deviations rate.

The second application deals with the error incurred when using the plug-in estimate $\min_{x \in D} \hat{f}_n(x)$ to predict the minimum value $\min_{x \in D} f(x)$. For this error, we prove the moderate deviations rate

$$P\left(\left|\min_{x \in D} \hat{f}_n(x) - \min_{x \in D} f(x)\right| \geq \delta\right) \lesssim \exp\left(-\delta^2 C' h_n^{-\frac{1}{2}s}\right). \tag{3.4}$$

Although (3.4) does not explicitly fix x, x^*, it can be obtained from the same theory because the mesh norm bound is uniform.

Both types of error probabilities are of broad interest in simulation, statistics, and uncertainty quantification. In particular, the pairwise comparison in (3.2) is motivated

by the approach developed by Glynn and Juneja [2004] for the ranking and selection problem, where one collects samples from a finite number of populations in an effort to select the one with the highest mean. The probability of correct selection can be related to the probability of false ordering between pairs of populations. The quantity $\pi_n(x, x^*)$ is the analog of this concept in the GP regression setting, with the additional complication that we are using a Bayesian model of f, so the event in (3.2) can only be viewed as an error conditionally given $f(x) \geq f(x^*)$.

Interestingly, the mesh norm is actually in use as a criterion for design of experiments, in the literature on so-called space-filling designs [Pronzato and Müller, 2012, Joseph et al., 2015]. As early as Johnson et al. [1990], statisticians have proposed to spread out the design points in D in a way that essentially minimizes the mesh norm. From (3.3), we can see that this has the effect of speeding up the rate at which (3.2) converges to zero, uniformly over all x, x^*. Essentially, if we have no specific x^* to serve as the reference solution, we can view space-filling designs as a way to minimize the probability of false ordering across *all* possible x^*.

The available theory for Gaussian process regression has extensively studied (pointwise) consistency; see, e.g., Ghosal and Roy [2006] or Bect et al. [2019]. With regard to convergence rate theory, our result is closest to the literature on design of experiments, where the design points are pre-selected; within this stream, Teckentrup [2020] and Wang et al. [2020] are two recent studies focusing on convergence rates for the estimation error of GP regression. Their object of study is different from the tail probabilities considered in our work, and so the rates have completely different orders, although their analysis also makes some connections to the mesh norm. A different, less directly related stream

of literature focuses on online optimization problems where the goal is to maximize the sum of the function values of the design points; a representative example of this type of work is Srinivas et al. [2012], with many subsequent developments focusing on algorithmic issues such as parallelization [Desautels et al., 2014]. In general, many of the existing rate results are derived for specific classes of kernels, such as squared exponential [Pati et al., 2015] and Matérn [Teckentrup, 2020, Vakili et al., 2020], or specific choices of the design points [Bull, 2011]. Some general tail probabilities were derived by Adler [2000] and Ghosal and Roy [2006], but they pertain to generic Gaussian processes, rather than the GP regression mechanism.

To our knowledge, this chapter presents the first moderate deviations results for Gaussian process regression estimators. It is well-known [Dembo and Zeitouni, 2009] that sample averages of i.i.d. Gaussian observations satisfy large deviations laws. Similar laws hold for ordinary least squares estimators under Gaussian residuals [Zhou and Ryzhov, 2021], extrema of Gaussian vectors [van der Hofstad and Honnappa, 2019], and various finite-dimensional statistical estimators [Arcones, 2006]. Gaussian process regression can be viewed as an infinite-dimensional generalization of linear regression, but the analysis is made much more complicated because, essentially, the dimensionality of the objects used to construct the estimator grows over time, and their asymptotic behavior heavily depends on the covariance kernel. One could perhaps recover large deviations laws for certain specific choices of the kernel and design, but it is far from clear whether this is possible in general. In the process of proving our results, we also establish a modified version of the Gärtner-Ellis theorem [Dembo and Zeitouni, 2009], which may be of stand-alone interest.

Section 3.2 describes the GP regression framework, states the assumptions used throughout this chapter, and gives important technical preliminaries. Section 3.3 gives the bulk of our analysis, which relies on a general large deviations law for random vectors. This latter result also requires some new technical developments, but since they are unrelated to GP regression, they are deferred to Section 3.5 for readability. Section 3.4 applies our analysis to derive (3.3) and (3.4), and presents several more explicit examples. Section 3.6 concludes.

3.2 Gaussian process regression and approximation theory

Section 3.2.1 presents some definitions, assumptions and properties pertaining to Gaussian process regression. Section 3.2.2 describes some important technical preliminaries from approximation theory.

3.2.1 Definitions and assumptions

Recalling the model in (3.1), we assume that the mean function m is Lipschitz continuous, and the Gaussian process $\mathcal{E}$ is specified by

$$\begin{aligned} \mathbb{E}\left(\mathcal{E}\left(x\right)\right) &= 0, \\ Cov\left(\mathcal{E}\left(x\right), \mathcal{E}\left(x'\right)\right) &= k\left(x, x'\right) \end{aligned}$$

for all $x, x' \in D$. We assume that $k : D \times D \to \mathbb{R}$ is a fixed, symmetric kernel function mapping $D \times D$ into $\mathbb{R}_+$. The kernel is required to be positive definite, meaning that, for

any n, any set of n distinct design points $\{x_m\}_{m=1}^n \subseteq D$, and any vector $v \in \mathbb{R}^n$, we have

$$\sum_{m,m'} v_m v_{m'} k(x_m, x_{m'}) > 0.$$

Without this assumption, the Gaussian process would be degenerate.

In addition, we assume that there exists a function ϕ on $\mathbb{R}_+$ such that $k(x, x') = \phi(\|x - x'\|)$ for all x, x'. Such a ϕ is called a *radial basis function*. We assume that ϕ is twice differentiable at zero with $\phi''(0) < 0$. Many commonly used covariance kernels satisfy this requirement, including Gaussian, multiquadric, inverse quadratic, inverse multiquadric and others.

Denote by $X_n = \{x_m\}_{m=1}^n$ the set of design points. We treat the design points as a deterministic sequence, as is standard in the literature on design of experiments, and assume that $\{x_n\}$ becomes dense in D as $n \to \infty$, a common condition in the theoretical literature [Vazquez and Bect, 2010b]. For convenience, we introduce the notation

$$\begin{aligned} f(X_n) &= (f(x_1), ..., f(x_n))^\top, \\ m(X_n) &= (m(x_1), ..., m(x_n))^\top, \\ K(X_n, x) &= (k(x, x_1), ..., k(x, x_n))^\top, \end{aligned}$$

as well as $K(x, X_n) = K(X_n, x)^\top$. We also denote by $K(X_n, X_n)$ the matrix whose (m, m')th entry is $k(x_m, x_{m'})$.

Given the design points X_n and observations $f(X_n)$, the posterior distribution of

$f(x)$, at any arbitrary $x \in D$, is Gaussian with mean

$$\hat{f}_n(x) = K(x, X_n) K(X_n, X_n)^{-1} f(X_n),$$

and variance

$$P_{X_n}(x) = k(x, x) - K(x, X_n) K(X_n, X_n)^{-1} K(X_n, x). \tag{3.5}$$

This specific structure of the mean and variance is what is referred to by the name of Gaussian process regression. The variance $P_{X_n}(x)$, viewed as a function of x, is also sometimes called the "power function" in the literature on interpolation. In this chapter, we use the posterior mean $\hat{f}_n$ to interpolate the observed function values $f(X_n)$ over the design space D and make predictions at unobserved points.

We let $\mathcal{H}$ denote the reproducing kernel Hilbert space (RKHS) whose reproducing kernel is k. The construction and uniqueness of $\mathcal{H}$ are discussed in Wendland [2004]. For our purposes, it is sufficient to review the following properties. Letting $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ be the inner product of $\mathcal{H}$, we know that:

1. $k(\cdot, x) \in \mathcal{H}$ for all $x \in D$.

2. $g(x) = \langle g, k(\cdot, x) \rangle_{\mathcal{H}}$ for all $g \in \mathcal{H}$ and $x \in D$.

3. $k(x, x') = \langle k(\cdot, x), k(\cdot, x') \rangle_{\mathcal{H}}$ for all $x, x' \in D$.

Additionally, from the usual properties of the inner product, we have the Cauchy-Schwarz inequality $|\langle g_1, g_2 \rangle_{\mathcal{H}}| \leq \|g_1\|_{\mathcal{H}} \|g_2\|_{\mathcal{H}}$, where $\|\cdot\|_{\mathcal{H}}$ is the norm induced by the inner

product.

3.2.2 Approximation theory

With the assumptions made in Section 3.2.1, Gaussian process regression can be seen as a special case of radial basis function (RBF) interpolation, enabling us to make use of some results from interpolation theory. We should note, however, that this theory treats interpolation models as purely deterministic, and thus has very different assumptions and interpretations than GP regression. Below, we present key facts from the theory that will be important for our analysis, and discuss their applicability to our setting when necessary.

Like GP regression, RBF interpolation requires a kernel k with the properties described in Section 3.2.1, as well as a matrix X_n describing n design points. Recall that, under these assumptions, we have $k(x, x') = \phi(\|x - x'\|)$. Denote by $\mathcal{L}_{k,X_n}$ the operator mapping some fixed function $g : D \to \mathbb{R}^d$ to its interpolant according to

$$\mathcal{L}_{k,X_n} g(x) = \sum_{m=1}^{n} \alpha_m k(x, x_m), \tag{3.6}$$

where the coefficients α_m solve the linear system

$$\sum_{m=1}^{n} \alpha_m k(x_m, x_{m'}) = g(x_{m'}), \qquad m' = 1, ..., n. \tag{3.7}$$

In fact, Wu and Schaback [1993] presents a more general form where (3.6)-(3.7) include additional polynomial functions, but this will not be necessary for our purposes. It can be shown that $\mathcal{L}_{k,X_n} g(x) = K(x, X_n) K(X_n, X_n)^{-1} g(X_n)$, similar to the calculations

used in GP regression.

Let $\tilde{g}$ be the Fourier transform of g, and suppose that the generalized Fourier transform of the function $x \mapsto \phi(\|x\|)$ exists and coincides with a continuous function $\tilde{\phi}$ on $\mathbb{R}^d \setminus \{0\}$ satisfying

$$0 < \tilde{\phi}(x) \leq c_{\tilde{\phi}} \|x\|^{-d-s_\infty} \qquad \text{as } \|x\| \to \infty \tag{3.8}$$

for suitable constants $c_{\tilde{\phi}}, s_\infty > 0$. Define

$$c_{g,\phi}^2 = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\tilde{g}(x)|^2 \, \tilde{\phi}(x)^{-1} \, dx.$$

The results below require $c_{g,\phi}^2 < \infty$, which essentially means that g resides in the RKHS whose reproducing kernel is k.

Before stating the key results, we should make it clear that we will *not* require $c_{f,\phi}^2 < \infty$, i.e., we will not apply the above definitions with f as the choice of g. It was shown in Lukić and Beder [2001] that a sample from a GP prior is almost surely not in the RKHS induced by the kernel assumed in the prior. Therefore, it is not possible for the function f to satisfy $c_{f,\phi}^2 < \infty$. This is a major difference between GP regression and interpolation theory, where f is modeled as a deterministic function and so the condition $c_{f,\phi}^2 < \infty$ is seen as fairly innocuous (for example, it is assumed in Wu and Schaback, 1993 and many other papers in pure interpolation theory, e.g., Li and Ryzhov, 2021). In the present work, however, we cannot make this assumption, and will instead apply this framework to *other* choices of g related to the kernel, for example the function $k(\cdot, x)$ for

fixed x.

For any compact $E \subseteq D$, let $h_n(E) = \max_{y \in E} \min_{m=1,\ldots,n} \|y - x_m\|_2$ be the mesh norm of E. We slightly abuse notation by using h_n to denote $h_n(D)$ when the entire domain is considered. Denote by $B_{x,\rho}$ the closed ball of radius ρ centered at $x \in \mathbb{R}^d$. We can now state the results that will be referenced and applied throughout this chapter.

Lemma 3.1 (Wu and Schaback, 1993)**.** *Fix $\rho > 0$ and assume that the kernel k satisfies (3.8) with some s_∞. Then, there exist positive constants $\bar{h}$ and c_P such that, for any X_n and any point $x \in \mathbb{R}^d$ with $h_n(B_{x,\rho} \cap D) < \bar{h}$, the power function P_{X_n} defined in (3.5) satisfies*

$$P_{X_n}(x) \leq c_P \left(h_n(B_{x,\rho} \cap D)\right)^{s_\infty}.$$

Lemma 3.2 (Wu and Schaback, 1993)**.** *Fix g satisfying $c_{g,\phi} < \infty$ and assume that the kernel k satisfies (3.8) with some s_∞. Then, for any X_n and any $x \in \mathbb{R}^d$, we have*

$$|g(x) - \mathcal{L}_{k,X_n} g(x)|^2 \leq c_{g,\phi}^2 P_{X_n}(x).$$

We note that, in Lemma 3.1, the constant c_P only depends on d and s_∞, but not on the fixed value ρ. The same is true of the ratio $\frac{\bar{h}}{\rho}$, indicating that $\bar{h}$ is proportional to ρ.

3.3 Large deviations for a fixed pair of points

We now fix $x, x^* \in D$ and focus on the sequence of random vectors

$$Z_n = \left(\hat{f}_n(x), \hat{f}_n(x^*), f(x), f(x^*)\right)^\top.$$

Letting μ_n be the probability law of Z_n, we will apply the inequality

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n (E) \leq - \inf_{u\in E} I(u), \tag{3.9}$$

which holds for any closed measurable set E and any sequence $\{a_n\}$ satisfying $\lim_{n\to\infty} a_n = \infty$. Our end goal is to characterize I and specify a_n and E in a manner that causes (3.9) to yield results such as (3.3) or (3.4).

Inequalities of the form (3.9) can be obtained by invoking the Gärtner-Ellis theorem from large deviations theory [Dembo and Zeitouni, 2009]. In general, the function I is derived in the following manner. First, denote by

$$\Psi_n (\gamma) = \log \mathbb{E}_{\mu_n} \left(e^{\langle \gamma, Z_n \rangle} \right)$$

the cumulant-generating function of Z_n. Here $\langle \cdot, \cdot \rangle$ is the usual L^2 inner product on $\mathbb{R}^p$. Then, define

$$\Psi (\gamma) = \limsup_{n\to\infty} \frac{1}{a_n} \Psi_n (a_n \gamma), \tag{3.10}$$

which is allowed to take values on the extended real number line (i.e., may be $+\infty$ for some γ). Then, I is obtained via the Fenchel-Legendre transform

$$I(u) = \sup_{\gamma\in\mathbb{R}^p} \{\langle \gamma, u \rangle - \Psi (\gamma)\} \tag{3.11}$$

of Ψ. Inequality (3.9) then follows under certain technical conditions on Ψ.

Unfortunately, the technical conditions of the classical Gärtner-Ellis theorem do

not hold in our setting, so additional analysis is required to obtain (3.9). This analysis is carried out in the setting of general random vectors, and thus is somewhat tangential to the setting of Gaussian process regression. For this reason, we defer it to Section 3.5 at the end of the chapter; the core result of that section is Theorem 3.14, which recovers the desired inequality. Here, we take (3.9) as given, referring readers to Theorem 3.14 for the proof, and focus on applying this inequality to the specific sequence $\{Z_n\}$.

Section 3.3.1 studies the cumulant-generating functions of $\{Z_n\}$ and characterizes Ψ. Section 3.3.2 then studies the Fenchel-Legendre transform of Ψ, and Section 3.3.3 analyzes the convergence rates of various terms that appear in the transform. Section 3.3.4 concludes the main moderate deviations inequality.

3.3.1 Analysis of cumulant-generating functions

We write Z_n as

$$\begin{bmatrix} \hat{f}_n(x) \\ \hat{f}_n(x^*) \\ f(x) \\ f(x^*) \end{bmatrix} = \begin{bmatrix} K(x, X_n) K(X_n, X_n)^{-1} f(X_n) \\ K(x^*, X_n) K(X_n, X_n)^{-1} f(X_n) \\ f(x) \\ f(x^*) \end{bmatrix}$$

$$= \begin{bmatrix} K(x, X_n) K(X_n, X_n)^{-1} & 0 & 0 \\ K(x^*, X_n) K(X_n, X_n)^{-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} f\left(\bar{X}_n\right),$$

where $\bar{X}_n = X_n \cup \{x, x^*\}$. The distribution of Z_n is Gaussian with mean vector $Am\left(\bar{X}_n\right)$ and covariance matrix $V_n = A\Sigma_n A^\top$, where

$$A = \begin{bmatrix} K(x, X_n) K(X_n, X_n)^{-1} & 0 & 0 \\ K(x^*, X_n) K(X_n, X_n)^{-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$\Sigma_n = \begin{bmatrix} K(X_n, X_n) & K(X_n, x) & K(X_n, x^*) \\ K(x, X_n) & k(x, x) & k(x, x^*) \\ k(x^*, X_n) & k(x^*, x) & k(x^*, x^*) \end{bmatrix}.$$

For convenience, we introduce the notation

$$\begin{aligned} Q_{X_n}(x) &= K(x, X_n) K(X_n, X_n)^{-1} K(X_n, x), \\ Q_{X_n}(x, x^*) &= K(x, X_n) K(X_n, X_n)^{-1} K(X_n, x^*). \end{aligned}$$

Then, the power function P_{X_n} in (3.5) can be written as $P_{X_n}(x) = k(x, x) - Q_{X_n}(x)$. We also use the analogous notation $P_{X_n}(x, x^*) = k(x, x^*) - Q_{X_n}(x, x^*)$. With some trivial computation, we obtain

$$V_n = \begin{bmatrix} Q_{X_n}(x) & Q_{X_n}(x, x^*) & Q_{X_n}(x) & Q_{X_n}(x, x^*) \\ Q_{X_n}(x, x^*) & Q_{X_n}(x^*) & Q_{X_n}(x, x^*) & Q_{X_n}(x^*) \\ Q_{X_n}(x) & Q_{X_n}(x, x^*) & k(x, x) & k(x, x^*) \\ Q_{X_n}(x, x^*) & Q_{X_n}(x^*) & k(x, x^*) & k(x^*, x^*) \end{bmatrix}.$$

Since Z_n follows a multivariate normal distribution, it straightforwardly follows that

$$\Psi_n(\gamma) = \gamma^\top Am\left(\bar{X}_n\right) + \frac{1}{2}\gamma^\top V_n \gamma$$

for any $\gamma \in \mathbb{R}^4$. Then, by (3.10),

$$\Psi(\gamma) = \gamma^\top \left(\lim_{n\to\infty} Am\left(\bar{X}_n\right)\right) + \frac{1}{2}\limsup_{n\to\infty} \gamma^\top (a_n V_n)\gamma, \tag{3.12}$$

provided that the limit on the right-hand side of (3.12) exists.

To study these limits, it is helpful to observe that $Q_{X_n}(x, x^*)$ can be viewed as the RBF interpolant of the function $k(x, \cdot)$ evaluated at the point x^*, or, equivalently, the RBF interpolant of $k(x^*, \cdot)$ evaluated at the point x. This allows us to leverage the results from approximation theory that were stated in Section 3.2.2.

First, we consider the limit of this 4 dimensional vector

$$\begin{aligned} Am\left(\bar{X}_n\right) = \Big[& K(x, X_n) K(X_n, X_n)^{-1} m(X_n) \\ & K(x^*, X_n) K(X_n, X_n)^{-1} m(X_n) \quad m(x) \quad m(x^*) \Big]. \end{aligned}$$

We may observe that $\mathcal{L}_{k,X_n} m(x)$ is a (differentiable) linear combination of the values $k(x, x_m)$. Hence, the difference $y \mapsto m(y) - \mathcal{L}_{k,X_n} m(y)$ is a Lipschitz function with scattered zeros, which are asymptotically dense around x and x^*. Consequently, $m(x) - \mathcal{L}_{k,X_n} m(x) \to 0$, whence $\lim_{n\to\infty} Am\left(\bar{X}_n\right) = m_0$, with

$$m_0 = \left(m(x), m(x^*), m(x), m(x^*)\right)^\top.$$

Thus, (3.12) becomes

$$\Psi(\gamma) = \gamma^\top m_0 + \frac{1}{2} \limsup_{n\to\infty} \gamma^\top (a_n V_n) \gamma. \tag{3.13}$$

The precise behavior of the limit superior will depend on a_n and the asymptotics of the matrix V_n.

3.3.2 Analysis of Fenchel-Legendre transform

We begin by examining the limit of V_n. It is easy to see that $P_{X_n}(y) \geq 0$ for all $y \in D$. Furthermore, by Lemma 3.1 we can see that $P_{X_n}(y) \to 0$ if the design points are asymptotically dense in D. This implies that $Q_{X_n}(x) \to k(x,x)$ and similarly $Q_{X_n}(x^*) \to k(x^*,x^*)$, with $k(x,x) = k(x^*,x^*)$ by the properties of the radial basis function. Although we do not know the sign of $P_{X_n}(x,x^*)$, we can note that

$$P_{X_n}(x,x^*) = k(x,x^*) - (\mathcal{L}_{k,X_n} k(\cdot,x^*))(x) = k(x^*,x) - (\mathcal{L}_{k,X_n} k(\cdot,x))(x^*).$$

By Lemma 3.2, we have $P_{X_n}(x,x^*)^2 \leq c^2_{k(\cdot,x^*),\phi} P_{X_n}(x)$. The finiteness of $c_{k(\cdot,x^*),\phi}$ can be verified. Therefore, if we are given an asymptotically dense design, we have $P_{X_n}(x,x^*) \to 0$, whence $Q_{X_n}(x,x^*) \to k(x,x^*)$. Thus, we have shown that $V_n \to V$

entrywise, where

$$V = \begin{bmatrix} k(x,x) & k(x,x^*) & k(x,x) & k(x,x^*) \\ k(x,x^*) & k(x,x) & k(x,x^*) & k(x,x) \\ k(x,x) & k(x,x^*) & k(x,x) & k(x,x^*) \\ k(x,x^*) & k(x,x) & k(x,x^*) & k(x,x) \end{bmatrix}.$$

It is easy to verify that V has eigenvalues

$$\lambda_1 = 2\left(k(x,x) + k(x,x^*)\right), \lambda_2 = 2\left(k(x,x) - k(x,x^*)\right)$$

with respective eigenvectors

$$U_1 = \frac{1}{2}(1,1,1,1)^\top, \qquad U_2 = \frac{1}{2}(1,-1,1,-1)^\top,$$

and $\lambda_3 = \lambda_4 = 0$ with respective eigenvectors

$$U_3 = \frac{1}{\sqrt{2}}(1,0,-1,0)^\top, \qquad U_4 = \frac{1}{\sqrt{2}}(0,1,0,-1)^\top. \tag{3.14}$$

Similarly, denote by $\lambda_{i,n}$ and $U_{i,n}$ (for $1 \leq i \leq 4$) the eigenvalues and corresponding eigenvectors of V_n. Since $V_n \to V$, we also have $\lambda_{i,n} \to \lambda_i$. Accordingly, we also have $U_{1,n} \to U_1$ and $U_{2,n} \to U_2$. However, the zero eigenvalue of V has multiplicity 2, so $U_{3,n}, U_{4,n}$ will converge to limits U_3', U_4' that belong to the span of U_3, U_4, but these limits

need not be U_3, U_4 themselves. We know, however, that

$$(U_3', U_4') = (U_3, U_4)\, T \tag{3.15}$$

where $T \in \mathbb{R}^{2\times 2}$ is an orthonormal matrix.

Looking back to (3.11) and (3.13), we can write the Fenchel-Legendre transform as

$$\begin{aligned}
I(u) &= \sup_{\gamma\in\mathbb{R}^4}\left\{(u-m_0)^\top\gamma - \frac{1}{2}\limsup_{n\to\infty}\gamma^\top(a_nV_n)\,\gamma\right\}\\
&= \sup_{\gamma\in\mathbb{R}^4}\left\{(u-m_0)^\top U\gamma - \frac{1}{2}\limsup_{n\to\infty}\gamma^\top U^\top(a_nV_n)\,U\gamma\right\}\\
&= \sup_{\gamma\in\mathbb{R}^4}\left\{(u-m_0)^\top U\gamma - \frac{1}{2}\limsup_{n\to\infty}a_n\gamma^\top U^\top U_n\Lambda_nU_n^\top U\gamma\right\}\\
&= \sup_{\gamma\in\mathbb{R}^4}\left\{(u-m_0)^\top U\gamma - \frac{1}{2}\limsup_{n\to\infty}\gamma^\top U^\top(a_nV_n)\,U\gamma\right\}\\
&= \sup_{\gamma\in\mathbb{R}^4}\left\{(u-m_0)^\top U\gamma - \frac{1}{2}\limsup_{n\to\infty}\sum_j\left(\sum_i\gamma_iU_i^\top U_{j,n}\right)^2a_n\lambda_{j,n}\right\}.
\end{aligned}$$

Observe that $\lim_{n\to\infty}U_i^\top U_{j,n} = 1_{\{i=j\}}$, whence

$$\limsup_{n\to\infty}\left(\sum_i\gamma_iU_i^\top U_{j,n}\right)^2a_n\lambda_{j,n} = \gamma_j^2\lambda_j\lim_{n\to\infty}a_n = \infty$$

as long as $\gamma_j \neq 0$ for $j\in\{1,2\}$. Therefore, the supremum in (3.11) can only be achieved

at γ for which $\gamma_1 = \gamma_2 = 0$, whence

$$
\begin{aligned}
& I(u) \qquad (3.16)\\
= & \sup_{\gamma_3,\gamma_4} \Big\{ (u-m_0)^\top U_3\gamma_3 + (u-m_0)^\top U_4\gamma_4 \\
& -\frac{1}{2}\limsup_{n\to\infty} \sum_j \left(\sum_i \gamma_i U_i^\top U_{j,n} \right)^2 a_n\lambda_{j,n} \Big\} \\
\geq & \sup_{\gamma_3,\gamma_4} \Bigg\{ (u-m_0)^\top U_3\gamma_3 + (u-m_0)^\top U_4\gamma_4 - \frac{1}{2}\sum_{j=1}^{2} \lambda_j \limsup_{n\to\infty} \left(\sum_{i=3}^{4} \gamma_i U_i^\top U_{j,n} \right)^2 a_n \\
& -\frac{1}{2}\sum_{j=3}^{4} \limsup_{n\to\infty} \left(\sum_{i=3}^{4} \gamma_i U_i^\top U_{j,n} \right)^2 \limsup_{n\to\infty} a_n\lambda_{j,n} \Bigg\} \\
= & \sup_{\gamma_3,\gamma_4} \Bigg\{ (u-m_0)^\top U_3\gamma_3 + (u-m_0)^\top U_4\gamma_4 - \frac{1}{2}\sum_{j=1}^{2} \lambda_j \limsup_{n\to\infty} \left(\sum_{i=3}^{4} \gamma_i U_i^\top U_{j,n} \right)^2 a_n \\
& -\frac{1}{2}\sum_{j=3}^{4} \left(T_{1,j-2}\gamma_3 + T_{2,j-2}\gamma_4 \right)^2 \limsup_{n\to\infty} a_n\lambda_{j,n} \Bigg\}. \qquad (3.17)
\end{aligned}
$$

The supremum value in (3.11) is thus governed by the rate at which a_n increases. If this rate is fast, we will have to take $\gamma_3 = \gamma_4 = 0$, leading to $I = 0$. To avoid this situation, a_n should be assigned the highest order that makes one of the limits superior in (3.17) finite. Some matrix perturbation analysis is required to understand the rate that a_n can take.

3.3.3 Perturbation analysis for rate function

Define the notation

$$\tilde{V} = V - V_n = \begin{bmatrix} P_{X_n}(x) & P_{X_n}(x,x^*) & P_{X_n}(x) & P_{X_n}(x,x^*) \\ P_{X_n}(x,x^*) & P_{X_n}(x^*) & P_{X_n}(x,x^*) & P_{X_n}(x^*) \\ P_{X_n}(x) & P_{X_n}(x,x^*) & 0 & 0 \\ P_{X_n}(x,x^*) & P_{X_n}(x^*) & 0 & 0 \end{bmatrix}.$$

Let us also write $U_{j,n} = \sum_i \nu_{ijn} U_i$. Then,

$$\begin{aligned} \lambda_{j,n} U_{j,n} &= \left(V - \tilde{V}\right) U_{j,n} \\ &= \sum_i \nu_{ijn} V U_i - \tilde{V} U_{j,n} \\ &= \sum_i \nu_{ijn} \lambda_i U_i - \tilde{V} U_{j,n}, \end{aligned}$$

where the last line follows because λ_i is an eigenvalue (and U_i is an eigenvector) of V. Left-multiplying by the unit vector U_i, we obtain

$$\nu_{ijn} \lambda_{j,n} = \nu_{ijn} \lambda_i - U_i^\top \tilde{V} U_{j,n}. \tag{3.18}$$

Recalling that $\lambda_3 = \lambda_4 = 0$ and $\lambda_{j,n} > 0$, we find that

$$\nu_{ijn} = -\frac{U_i^\top \tilde{V} U_{j,n}}{\lambda_{j,n}}, \quad i \in \{3,4\}, \; j \in \{1,2\}. \tag{3.19}$$

This allows us to bound the limits superior in (3.17) as shown in Lemmas 3.3 and 3.4 below.

Lemma 3.3. *For fixed $\gamma_3, \gamma_4 \in \mathbb{R}$, we have*

$$\left(\sum_{i=3}^{4} \gamma_i U_i^\top U_{j,n}\right)^2 = O\left(P_{X_n}^2(x) + P_{X_n}^2(x, x^*) + P_{X_n}^2(x^*)\right), \quad j \in \{1, 2\}.$$

Proof: Using (3.19), we write

$$\left(\sum_{i=3}^{4} \gamma_i U_i^\top U_{j,n}\right)^2 = \left(\nu_{3jn}\gamma_3 + \nu_{4jn}\gamma_4\right)^2 = \frac{\left(U_{j,n}^\top \tilde{V}\left(\gamma_3 U_3 + \gamma_4 U_4\right)\right)^2}{\lambda_{j,n}^2}.$$

Plugging in the closed-form expressions for U_3, U_4 from (3.14) yields

$$\tilde{V}U_3 = \frac{1}{\sqrt{2}}\left(0, 0, P_{X_n}(x), P_{X_n}(x, x^*)\right)^\top, \qquad \tilde{V}U_4 = \frac{1}{\sqrt{2}}\left(0, 0, P_{X_n}(x, x^*), P_{X_n}(x^*)\right)^\top.$$

Since γ_3, γ_4 are fixed and U_3, U_4 are unit vectors, the Cauchy-Schwarz inequality yields

$$\left(\sum_{i=3}^{4} \gamma_i U_i^\top U_{j,n}\right)^2 \leq \frac{\max\left\{\gamma_3^2, \gamma_4^2\right\}}{2\lambda_{j,n}^2}\left(P_{X_n}^2(x) + P_{X_n}^2(x, x^*) + P_{X_n}^2(x^*)\right)$$

as desired. □

Lemma 3.4. *Suppose that the matrix T defined in (3.15) has no zero-valued entries. Then,*

$$\begin{aligned}\lambda_{j,n} &= \left(\frac{1}{2} + o(1)\right)\left(P_{X_n}(x) + \frac{T_{2,j-2}}{T_{1,j-2}} P_{X_n}(x, x^*)\right) \\ &= \left(\frac{1}{2} + o(1)\right)\left(P_{X_n}(x^*) + \frac{T_{1,j-2}}{T_{2,j-2}} P_{X_n}(x, x^*)\right). \end{aligned} \tag{3.20}$$

for $j \in \{3,4\}$.

Proof: Recall (3.18) and note that $\nu_{ijn} \to T_{i-2,j-2}$ for $i,j \in \{3,4\}$. Since T is assumed to have no zero-valued entries, we do not need to worry about zero values of ν_{ijn}. Then, (3.19) can be rewritten as

$$\lambda_{j,n} = -\frac{U_i^\top \tilde{V} U_{j,n}}{\nu_{ijn}}, \quad i,j \in \{3,4\}\,. \tag{3.21}$$

The first equality in (3.20) can be obtained by setting $i = 3$, whence (3.21) yields

$$\lambda_{j,n} = -\frac{1}{\nu_{3jn}\sqrt{2}} \left(0, 0, P_{X_n}(x), P_{X_n}(x, x^*)\right) \cdot U_{j,n}.$$

By expressing $(0,0,1,0)$ and $(0,0,0,1)$ in terms of U_i, we obtain

$$\begin{aligned}
\lambda_{j,n} &= -\frac{1}{\nu_{3jn}\sqrt{2}} \left(P_{X_n}(x)\, U \cdot \left(\frac{1}{2}, \frac{1}{2}, -\frac{1}{\sqrt{2}}, 0\right)^\top \right. \\
&\quad \left. + P_{X_n}(x, x^*)\, U \cdot \left(\frac{1}{2}, -\frac{1}{2}, 0, -\frac{1}{\sqrt{2}}\right)^\top \right)^\top U_{j,n} \\
&= -\frac{1}{\nu_{3jn}\sqrt{2}} \left(P_{X_n}(x) \cdot \left(\frac{1}{2}, \frac{1}{2}, -\frac{1}{\sqrt{2}}, 0\right) \right. \\
&\quad \left. + P_{X_n}(x, x^*) \cdot \left(\frac{1}{2}, -\frac{1}{2}, 0, -\frac{1}{\sqrt{2}}\right) \right) v_{\cdot jn} \\
&= \left(\frac{1}{2} + o(1)\right) \left(P_{X_n}(x) + \frac{T_{2,j-2}}{T_{1,j-2}} P_{X_n}(x, x^*) \right),
\end{aligned}$$

where the last line follows from the fact that $\nu_{ijn} \to 0$ for $i \in \{1,2\}$ and $j \in \{3,4\}$, while $\nu_{ijn} \to T_{i-2,j-2}$ for $i,j \in \{3,4\}$. The second equality in (3.20) can be obtained by repeating the above arguments with $i = 4$. □

The analysis in Lemma 3.4 is easily extended to handle situations where T has zero-valued entries. If this occurs, we must have either $T_{11} = T_{22} = 0$ or $T_{12} = T_{21} = 0$ because T is orthonormal. In the first case, we can repeat the proof of Lemma 3.4 with $i = 4, j = 3$ and $i = 3, j = 4$ and obtain

$$\lambda_{3,n} = \left(\frac{1}{2} + o(1)\right) P_{X_n}(x^*), \qquad \lambda_{4,n} = \left(\frac{1}{2} + o(1)\right) P_{X_n}(x). \tag{3.22}$$

In the second case, we repeat the same proof with $i = 3, j = 3$ and $i = 4, j = 4$ and obtain

$$\lambda_{3,n} = \left(\frac{1}{2} + o(1)\right) P_{X_n}(x), \qquad \lambda_{4,n} = \left(\frac{1}{2} + o(1)\right) P_{X_n}(x^*).$$

In general, the bounds in Lemmas 3.3 and 3.4 depend on $P_{X_n}(x, x^*)$, which is a difficult object to study. Lemma 3.5 establishes a bound that relates this quantity to a simpler function of the design points. Then, Lemma 3.6 derives a similar lower bound on $P_{X_n}(x)$. Note that these results provide *lower* bounds; we will later multiply them by negative quantities to convert them to upper bounds, which will enable additional analysis of the terms in (3.17).

Lemma 3.5. *Let $\lambda_{\min}(\cdot)$ denote the smallest eigenvalue of a square matrix. The following bound holds:*

$$2P_{X_n}(x, x^*) + P_{X_n}(x) + P_{X_n}(x^*) \geq 2\lambda_{\min}\left(K\left(\bar{X}_n, \bar{X}_n\right)\right).$$

Proof: For notational convenience, define a vector $\kappa(x) = K(x, X_n) K(X_n, X_n)^{-1}$.

Note that $\kappa(x)$ takes values in $\mathbb{R}^n$, and observe the identities

$$\begin{aligned}\sum_{m=1}^{n}\left(\kappa_m(x)+\kappa_m(x^*)\right)k(x_m,x) &= Q_{X_n}(x)+Q_{X_n}(x,x^*)\\ \sum_{m=1}^{n}\left(\kappa_m(x)+\kappa_m(x^*)\right)k(x_m,x^*) &= Q_{X_n}(x^*)+Q_{X_n}(x,x^*)\end{aligned}$$

and

$$\begin{aligned}&\sum_{m,m'}\left(\kappa_m(x)+\kappa_m(x^*)\right)\left(\kappa_{m'}(x)+\kappa_{m'}(x^*)\right)k(x_m,x_{m'})\\ &=Q_{X_n}(x)+2Q_{X_n}(x,x^*)+Q_{X_n}(x^*).\end{aligned}$$

We extend κ to $\mathbb{R}^{n+2}$ by taking $\kappa_{n+1},\kappa_{n+2}\equiv-\frac{1}{2}$. Plugging in the above identities, we derive

$$\begin{aligned}&\sum_{m,m'=1}^{n+2}\left(\kappa_m(x)+\kappa_m(x^*)\right)\left(\kappa_{m'}(x)+\kappa_{m'}(x^*)\right)k(x_m,x_{m'})\\ =\;&\left(\kappa_{n+1}(x)+\kappa_{n+1}(x^*)\right)^2k(x,x)+\left(\kappa_{n+2}(x)+\kappa_{n+2}(x^*)\right)^2k(x^*,x^*)\\ &+2\left(\kappa_{n+1}(x)+\kappa_{n+1}(x^*)\right)\left(\kappa_{n+2}(x)+\kappa_{n+2}(x^*)\right)k(x,x^*)\\ &+2\left(\kappa_{n+1}(x)+\kappa_{n+1}(x^*)\right)\left(Q_X(x)+Q_X(x,x^*)\right)\\ &+2\left(\kappa_{n+2}(x)+\kappa_{n+2}(x^*)\right)\left(Q_X(x^*)+Q_X(x,x^*)\right)\\ &+Q_{X_n}(x)+2Q_{X_n}(x,x^*)+Q_{X_n}(x^*)\\ =\;&2P_{X_n}(x,x^*)+P_{X_n}(x)+P_{X_n}(x^*).\end{aligned}$$

Thus, we arrive at

$$\begin{aligned}
& 2P_{X_n}(x, x^*) + P_{X_n}(x) + P_{X_n}(x^*) \\
= \; & (\kappa(x) + \kappa(x^*))^\top K\left(\bar{X}_n, \bar{X}_n\right)(\kappa(x) + \kappa(x^*)) \\
\geq \; & \|\kappa(x) + \kappa(x^*)\|_2^2 \cdot \lambda_{\min}\left(K\left(\bar{X}_n, \bar{X}_n\right)\right) \\
= \; & \left(\sum_{m=1}^{n} (\kappa_m(x) + \kappa_m(x^*))^2 + 2\right) \lambda_{\min}\left(K\left(\bar{X}_n, \bar{X}_n\right)\right) \\
\geq \; & 2\lambda_{\min}\left(K\left(\bar{X}_n, \bar{X}_n\right)\right),
\end{aligned}$$

which completes the proof. □

Lemma 3.6. *Let* $X_n' = X_n \cup \{x\}$. *Then,*

$$P_{X_n}(x) \geq \lambda_{\min}\left(K\left(X_n', X_n'\right)\right).$$

Proof: Define $\kappa(x)$ as in the proof of Lemma 3.5 and extend it to $\mathbb{R}^{n+1}$ by taking $\kappa_{n+1} \equiv -1$. Then, by repeating the arguments in the proof of Lemma 3.5, we obtain

$$\begin{aligned}
P_{X_n}(x) &= \sum_{m,m'=1}^{n+1} \kappa_m(x)\,\kappa_{m'}(x)\,k(x_m, x_{m'}) \\
&\geq \lambda_{\min}\left(K\left(X_n', X_n'\right)\right) \sum_{m=1}^{n+1} \kappa_m^2(x) \\
&\geq \lambda_{\min}\left(K\left(X_n', X_n'\right)\right),
\end{aligned}$$

as desired. □

Now, we can study the rate at which $\lambda_{\min}\left(K\left(X_n', X_n'\right)\right)$ or $\lambda_{\min}\left(K\left(\bar{X}_n, \bar{X}_n\right)\right)$ con-

verges to zero. For this, we cite the following result (Theorem 12.3 of Wendland, 2004).

Lemma 3.7 (Wendland, 2004)**.** *Define* $q_{X_n} = \min_{x_m \neq x_{m'}} \|x_m - x_{m'}\|_2$ *and* $\phi_0(M) = \inf_{\|y\|_2 \leq M} \tilde{\phi}(y)$*, where* $\tilde{\phi}$ *is the generalized Fourier transform of the radial basis function* ϕ*. Then,*

$$\lambda_{\min}\left(K\left(X_n, X_n\right)\right) \geq C_d \phi_0\left(\frac{M_d}{q_{X_n}}\right) q_{X_n}^{-d},$$

where the constants C_d, M_d *depend only on* d*.*

Combining Lemmas 3.5-3.7, we have

$$P_{X_n}(x) \geq C_d \phi_0\left(\frac{M_d}{q_{X_n'}}\right) q_{X_n'}^{-d}.$$

Consequently, the inequality in Lemma 3.5 becomes

$$2P_{X_n}(x, x^*) + P_{X_n}(x) + P_{X_n}(x^*) \geq 2C_d \phi_0\left(\frac{M_d}{q_{\bar{X}_n}}\right) q_{\bar{X}_n}^{-d}. \tag{3.23}$$

The lower bound in (3.23) can be connected back to the upper bound obtained in Lemma 3.4 in the following manner. Note that, if T has no zero-valued entries as assumed in Lemma 3.4, by orthogonality we either have $\frac{T_{11}}{T_{21}} > 0$ and $\frac{T_{12}}{T_{22}} < 0$, or vice versa (note also that $T_{11} = -T_{22}$). Without loss of generality, we only treat the first case here.

Supposing that $\frac{T_{12}}{T_{22}} < 0$, we apply (3.23) to (3.20) with $j = 4$ and argue

$$\lambda_{4,n} \leq \left(\frac{1}{2} + o(1)\right)\left[P_{X_n}(x) - \frac{T_{22}}{2T_{12}}\left(P_{X_n}(x) + P_{X_n}(x^*) - 2C_d\phi_0\left(\frac{M_d}{q_{\bar{X}_n}}\right)q_{\bar{X}_n}^{-d}\right)\right] \tag{3.24}$$

$$= \left(\frac{1}{2} + o(1)\right)\left[\left(1 - \frac{T_{22}}{2T_{12}}\right)P_{X_n}(x) - \frac{T_{22}}{2T_{12}}P_{X_n}(x^*) + O\left(q_{\bar{X}_n}^{s_\infty}\right)\right], \tag{3.25}$$

$$= O\left(h_n\left(B_{x,\rho} \cap D\right)^{s_\infty}\right). \tag{3.26}$$

In this derivation, (3.24) uses the fact that $\frac{T_{12}}{T_{22}} < 0$ to convert the lower bound in (3.23) into an upper bound, while (3.25) applies (3.8) to bound ϕ_0. Noting that the multipliers $1 - \frac{T_{22}}{2T_{12}}$ and $-\frac{T_{22}}{2T_{12}}$ are both strictly positive, we then obtain (3.26) by applying Lemma 3.1 together with the fact that

$$h_n\left(B_{x,\rho} \cap D\right) \geq q_{X_n} \geq q_{\bar{X}_n}.$$

Next, we return to (3.20) with $j = 3$ and obtain

$$\lambda_{3,n} \leq \left(\frac{1}{2} + o(1)\right)\left(P_{X_n}(x) + \frac{T_{21}}{T_{11}}\sqrt{\phi(0)}\sqrt{P_{X_n}(x)}\right)$$

by using the Cauchy-Schwarz inequality for the RKHS inner product to produce the simple bound

$$|P_{X_n}(x, x^*)| \leq \sqrt{\phi(0)}\sqrt{P_{X_n}(x)}.$$

Applying Lemma 3.1, we conclude that $\lambda_{3,n} = O\left(h_n\left(B_{x,\rho} \cap D\right)^{\frac{1}{2}s_\infty}\right)$. Finally, applying

Lemma 3.1 to the bound in Lemma 3.3 straightforwardly yields

$$\left(\sum_{i=3}^{4}\gamma_i U_i^\top U_{j,n}\right)^2 = O\left(h_n\left(B_{x,\rho}\cap D\right)^{s_\infty}\right).$$

When the design points are asymptotically dense in D, we have $h_n\left(B_{x,\rho}\cap D\right)\leq h_n\left(D\right)$ with $h_n\left(D\right)\to 0$. Thus, among the limits superior in (3.17), one is $O\left(h_n^{\frac{1}{2}s_\infty}\right)$, and the others are $O\left(h_n^{s_\infty}\right)$. This will also happen in the symmetric situation where $\frac{T_{12}}{T_{22}}>0$, but with the order switched for $\lambda_{3,n}$ and $\lambda_{4,n}$.

3.3.4 Main moderate deviations inequality

The conclusions of Section 3.3.3 suggest that a_n should have the exact order $h_n^{-\frac{1}{2}s_\infty}$, which we denote by $a_n\sim h_n^{-\frac{1}{2}s_\infty}$. Then, we obtain

$$\left(\sum_{i=3}^{4}\gamma_i U_i^\top U_{j,n}\right)^2 a_n\to 0$$

in (3.17), and bound $I\left(u\right)\geq I^l\left(u\right)$, where I^l is defined as

$$I^l\left(u\right)=\sup_{\gamma_3,\gamma_4\in\mathbb{R}}\left\{\left(u-m_0\right)^\top U_3\gamma_3+\left(u-m_0\right)^\top U_4\gamma_4-\frac{1}{2}c_3\left(T_{11}\gamma_3+T_{21}\gamma_4\right)^2\right\}$$

when $\frac{T_{12}}{T_{22}}<0$, and

$$I^l\left(u\right)=\sup_{\gamma_3,\gamma_4\in\mathbb{R}}\left\{\left(u-m_0\right)^\top U_3\gamma_3+\left(u-m_0\right)^\top U_4\gamma_4-\frac{1}{2}c_4\left(T_{12}\gamma_3+T_{22}\gamma_4\right)^2\right\}$$

when $\frac{T_{12}}{T_{22}} > 0$, for some suitable constants c_3, c_4. Furthermore, recalling (3.14) and the definition of m_0, we find that $m_0^\top U_3 = m_0^\top U_4 = 0$. Now, applying (3.9), we can finally state our main result.

Theorem 3.8. *Let T be as in (3.15), take $a_n \sim h_n^{-\frac{1}{2}s_\infty}$ and let c_l be a constant satisfying $\limsup_{n\to\infty} a_n\lambda_{j,n} \leq c_l$ for $j \in \{3,4\}$. If $||T_{11}| - |T_{21}|| \neq 1$, we have*

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(E) \leq -\inf_{u\in E} I^l(u) \tag{3.27}$$

for any closed $E \subseteq D$, with

$$I^l(u) = \sup_{\gamma_3,\gamma_4\in\mathbb{R}} \left\{ u^\top U_3\gamma_3 + u^\top U_4\gamma_4 - \frac{1}{2}c_l \left(|T_{11}|\gamma_3 + |T_{21}|\gamma_4\right)^2 \right\}. \tag{3.28}$$

Throughout this analysis, we have assumed that the design is asymptotically dense, but it is possible to recover Theorem 4.10, for fixed x, x^*, as long as the design is dense only in neighborhoods of those two points, e.g., in $B_{x,\rho} \cup B_{x^*,\rho}$ for some $\rho > 0$. In that case a_n will take the order of $\min\left\{h_n(B_{x,\rho})^{-\frac{1}{2}s_\infty}, h_n(B_{x^*,\rho})^{-\frac{1}{2}s_\infty}\right\}$.

Finally, we note that the right-hand side of (3.27) is some strictly negative, problem-specific constant, and it is the order of a_n that governs the convergence rate of $\mu_n(E)$. Our result shows how the rate depends on the kernel through the quantity s_∞. Examples of various kernels and their s_∞ values can be found in Wu and Schaback [1993]. Note that the rate function I^l has no dependence on the mean function m of the Gaussian process model.

3.4 Applications: pairwise comparisons and estimation error

Sections 3.4.1-3.4.2 apply Theorem 4.10 to prove (3.3) and (3.4), respectively. It is interesting to note that the proofs are overall very similar, but use different definitions of the error set E in (3.27). This illustrates the flexibility of our framework, as one can obtain very different types of inequalities simply by changing the error set. Section 3.4.3 presents several other results of interest where the moderate deviations bound can be made more explicit.

3.4.1 Moderate deviations for false ordering

We return to (3.2) and write

$$\pi_n\left(x, x^*\right) = \frac{P\left(\hat{f}_n\left(x\right) \leq \hat{f}_n\left(x^*\right) - \delta,\ f\left(x\right) \geq f\left(x^*\right)\right)}{P\left(f\left(x\right) \geq f\left(x^*\right)\right)}. \tag{3.29}$$

For fixed x, x^*, the denominator is a strictly positive constant, so we can focus on the numerator, which fits into the framework of Section 3.3 with

$$E = \left\{u \in \mathbb{R}^4 \,:\, u_1 \leq u_2 - \delta,\ u_3 \geq u_4\right\}. \tag{3.30}$$

We will apply Theorem 4.10 and derive a more explicit form for (3.28). First, note that the supremum in (3.28) can only be finite when

$$\frac{u^\top U_3}{|T_{11}|} = \frac{u^\top U_4}{|T_{21}|}. \tag{3.31}$$

Letting η be the value in (3.31), we then have $I^l(u) = \frac{\eta^2}{2c_l}$. Then, we minimize $I^l(u)$ subject to (3.30)-(3.31). From the optimality conditions, it can be seen that the inequalities in (3.30) must be binding at optimality, which leads to

$$\inf_{u \in E} I^l(u) = \frac{\delta^2}{4c_l} \frac{1}{(|T_{11}| - |T_{21}|)^2}.$$

Applying Theorem 4.10, we complete the proof of (3.3). The formal statement of the result is as follows.

Theorem 3.9. *Let T be as in (3.15), take $a_n = O\left(h_n^{-\frac{1}{2}s_\infty}\right)$ and let c_l be a constant satisfying $\limsup_{n\to\infty} a_n \lambda_{j,n} \leq c_l$ for $j \in \{3,4\}$. If $||T_{11}| - |T_{21}|| \notin \{0,1\}$, we have*

$$\pi_n(x, x^*) \leq C_1 \exp\left(-\frac{\delta^2 C_2}{4c_l (|T_{11}| - |T_{21}|)^2} h_n^{-\frac{1}{2}s_\infty}\right)$$

where C_1, C_2 are positive constants.

In fact, we can show that the moderate deviations bound holds uniformly for all $x, x^* \in D$. To do so, we must make sure that the denominator of (3.29) is well-behaved. It is easily seen that

$$P(f(x) \geq f(x^*)) = \Phi\left(\frac{m(x) - m(x^*)}{\sqrt{2(k(x,x) - k(x,x^*))}}\right),$$

where Φ is the standard Gaussian cdf. We let c_L be the Lipschitz constant of m, and

derive

$$\begin{aligned}
\lim_{\|x-x^*\|\to 0} \frac{\left(m\left(x\right)-m\left(x^*\right)\right)^2}{2\left(k\left(x,x\right)-k\left(x,x^*\right)\right)} &\leq \lim_{\|x-x^*\|\to 0} \frac{c_L^2\|x-x^*\|_2^2}{2\left(\phi\left(0\right)-\phi\left(\|x-x^*\|\right)\right)} \\
&= \frac{c_L^2}{2}\lim_{y\searrow 0}\frac{y^2}{\phi\left(0\right)-\phi\left(y\right)} \\
&= -c_L^2\lim_{y\searrow 0}\frac{y}{\phi'\left(y\right)} \\
&= -\frac{c_L^2}{\phi''\left(0\right)} \\
&< \infty
\end{aligned}$$

using the assumption made in Section 3.2.1 that ϕ is twice differentiable at zero with $\phi''\left(0\right)<0$. Because D is compact, there exists some $c_D>0$ satisfying

$$\inf_{x,x^*\in D} P\left(f\left(x\right)\geq f\left(x^*\right)\right)\geq c_D.$$

Furthermore, the constant C_2 in Theorem 3.9 does not depend on x, x^*. The constant C_1 may depend on x, x^*, but we can take C_1' to be its largest value over the compact set D. We then conclude the following.

Corollary 3.10. *Suppose that we are in the situation of Theorem 3.9. Then,*

$$\sup_{x,x^*\in D}\pi_n\left(x,x^*\right)\leq\frac{C_1'}{c_D}\exp\left(-\frac{\delta^2 C_2}{4c_l\left(\left|T_{11}\right|-\left|T_{21}\right|\right)^2}h_n^{-\frac{1}{2}s_\infty}\right)$$

where C_1', C_2, c_D are positive constants.

Finally, we consider two special cases not covered by Theorem 3.9. First, in the

case where $T_{11} = T_{21}$ or $T_{12} = T_{22}$, we straightforwardly obtain

$$\pi_n(x, x^*) \leq C_1 \exp\left(-\delta^2 C_s h_n^{-\frac{1}{2}s_\infty}\right)$$

for any $C_s > 0$.

The second and more important special case arises when $||T_{11}| - |T_{21}|| = 1$, i.e.,

$$T \in \left\{ \begin{bmatrix} \pm 1 & 0 \\ 0 & \pm 1 \end{bmatrix}, \begin{bmatrix} 0 & \pm 1 \\ \pm 1 & 0 \end{bmatrix} \right\}$$

We only present the first case, as the second is handled symmetrically. Recall that, in this situation, the rate behavior of $\lambda_{3,n}, \lambda_{4,n}$ is described by (3.22). Repeating the analysis of Sections 3.3.2-3.3.3, we can take $a_n \sim h_n^{-s_\infty+\varepsilon}$ for any $\varepsilon > 0$. Then, all of the limits superior in (3.17) vanish to zero, yielding

$$I(u) = \sup_{\gamma_3,\gamma_4 \in \mathbb{R}} \left\{ u^\top U_3 \gamma_3 + u^\top U_4 \gamma_4 \right\} = \begin{cases} 0 & u^\top U_3 = u^\top U_4 = 0, \\ \infty & \text{otherwise.} \end{cases}$$

However, the error set E does not contain any u that would satisfy $u^\top U_3 = u^\top U_4 = 0$, so $\inf_{u \in E} I(u) = \infty$. Consequently, we may conclude that

$$\pi_n(x, x^*) = o\left(\exp\left(-\delta^2 C_s h_n^{-s_\infty+\varepsilon}\right)\right), \quad \forall \varepsilon > 0.$$

3.4.2 Moderate deviations for estimation error

Let $\hat{x}_n = \arg\min_{x \in D} \hat{f}_n(x)$ and $\bar{x} = \arg\min_{x \in D} f(x)$. We will study the convergence rate of the tail probability $P\left(\left|\hat{f}_n(\hat{x}_n) - f(\bar{x})\right| \geq \delta\right)$ of the estimation error.

First, we write

$$\begin{aligned} P\left(\hat{f}_n(\hat{x}_n) \leq f(\bar{x}) - \delta\right) &= P\left(\hat{f}_n(\hat{x}_n) \leq f(\bar{x}) - \delta,\ f(\hat{x}_n) \geq f(\bar{x})\right) \\ &\leq \sup_{x, x^* \in D} P\left(\hat{f}_n(x) \leq f(x^*) - \delta,\ f(x) \geq f(x^*)\right), \end{aligned}$$

where the first line uses the fact that $f(\hat{x}_n) \geq f(\bar{x})$ by the definition of $\bar{x}$. Now, we may obtain a rate for $P\left(\hat{f}_n(\hat{x}_n) \leq f(\bar{x}) - \delta\right)$ by repeating the analysis of Section 3.4.1, but with (3.30) replaced by

$$E' = \left\{u \in \mathbb{R}^4 \,:\, u_1 \leq u_4 - \delta,\ u_3 \geq u_4\right\}. \tag{3.32}$$

Minimizing $I^l(u)$ subject to (3.31) and (3.32), we find that

$$\inf_{u \in E'} I^l(u) = \frac{\delta^2}{4 c_l T_{11}^2}.$$

For the other side of the error event, we write

$$\begin{aligned} P\left(\hat{f}_n(\hat{x}_n) \geq f(\bar{x}) + \delta\right) &= P\left(\hat{f}_n(\hat{x}_n) \geq f(\bar{x}) + \delta,\ \hat{f}_n(\bar{x}) \geq \hat{f}_n(\hat{x}_n)\right) \\ &\leq \sup_{x, x^* \in D} P\left(\hat{f}_n(x) \geq f(x^*) + \delta,\ \hat{f}_n(x^*) \geq \hat{f}_n(x)\right). \end{aligned}$$

Again, we repeat the analysis of Section 3.4.1, but with (3.30) replaced by

$$E'' = \left\{u \in \mathbb{R}^4 \,:\, u_1 \geq u_4 + \delta,\, u_2 \geq u_1\right\}. \tag{3.33}$$

Minimizing $I^l(u)$ subject to (3.31) and (3.33), we find that

$$\inf_{u \in E'} I^l(u) = \frac{\delta^2}{4c_l T_{21}^2}.$$

We then combine the preceding results with the arguments of Corollary 3.10 to complete the proof of (3.4). The final result is formally stated as follows. The situation where $||T_{11}| - |T_{21}|| = 1$ can be handled using the same arguments that were presented in Section 3.4.1.

Theorem 3.11. *Let T be as in (3.15), take $a_n = O\left(h_n^{-\frac{1}{2}s_\infty}\right)$ and let c_l be a constant satisfying $\limsup_{n\to\infty} a_n \lambda_{j,n} \leq c_l$ for $j \in \{3,4\}$. If $||T_{11}| - |T_{21}|| \neq 1$, we have*

$$P\left(\left|\min_{x\in D} \hat{f}_n(x) - \min_{x\in D} f(x)\right| \geq \delta\right) \leq 2C_1' \exp\left(-\frac{\delta^2 C_2}{4c_l \max\left\{T_{11}^2, T_{21}^2\right\}^2} h_n^{-\frac{1}{2}s_\infty}\right)$$

where C_1', C_2 are positive constants.

3.4.3 Other results of interest

In the following, we give several examples in which our main results can be made more explicit. To avoid excessive repetition, we focus on the uniform bound in Corollary 3.10 in our presentation, but analogs of the other results in Sections 3.4.1-3.4.2 can be

straightforwardly obtained as well. For simplicity, let us take $D = [0,1]^d$.

Gaussian kernel. Suppose that k is the Gaussian kernel with parameter α, that is, $k(x, x^*) = \exp(-\alpha\|x - x^*\|_2^2)$. For this particular kernel, it is known that s_∞ can take arbitrarily large values. However, Theorem 11.22 of Wendland [2004] proves the bound

$$P_{X_n}(x) \leq \exp\left(c_\alpha \frac{\log h_n}{h_n}\right),$$

where c_α depends only on α, d and D. In addition, Corollary 12.4 in Wendland [2004] provides a modified version of Lemma 3.7 for this setting, namely,

$$\lambda_{\min}(K(X_n, X_n)) \geq c'_\alpha \exp\left(-40.71 \frac{d^2}{\alpha q_{X_n}^2}\right) q_{X_n}^{-d}.$$

Thus, using the above results instead of Lemmas 3.1 and 3.7, we can repeat our analysis with $a_n \sim \exp\left(-c_\alpha \frac{\log h_n}{h_n}\right)$ and obtain, e.g.,

$$\sup_{x,x^* \in D} \pi_n(x, x^*) \leq c_1 \exp\left(-\delta^2 c_2 \exp\left(-\frac{c_\alpha}{2} \frac{\log h_n}{h_n}\right)\right)$$

under the same assumptions as Corollary 3.10.

Uniform design. Consider a uniform grid, discretized evenly in each dimension, with n being the total number of points in the discretization. One can find that $h_n = O\left(n^{-\frac{1}{d}}\right)$, leading to the explicit rate

$$\sup_{x,x^* \in D} \pi_n(x, x^*) \leq \frac{C'_1}{c_D} \exp\left(-\frac{\delta^2 C_2}{4c_l \left(|T_{11}| - |T_{21}|\right)^2} n^{\frac{1}{2d} s_\infty}\right)$$

under the assumptions of Corollary 3.10.

Uniform random design. Suppose that the design points are sampled from a uniform distribution on $[0,1]^d$. By adapting results in Janson [1987], one can show that $h_n = O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{d}}\right)$, leading to the explicit rate

$$\sup_{x,x^* \in D} \pi_n(x, x^*) \leq \frac{C_1'}{c_D} \exp\left(-\frac{\delta^2 C_2}{4c_l\left(|T_{11}| - |T_{21}|\right)^2}\left(\frac{n}{\log n}\right)^{\frac{1}{2d} s_\infty}\right)$$

under the assumptions of Corollary 3.10. One can also extend this result to a setting with independent, but non-uniform sampling. Suppose that the nth design point is sampled independently from some fixed density g_n with support $[0,1]^d$. Then, one can show that

$$h_n = O\left(\left(\frac{\log(c_g n)}{c_g n}\right)^{\frac{1}{d}}\right),$$

and the rate follows.

We remark that the above discussion implicitly assumes that $||T_{11}| - |T_{21}|| \notin \{0,1\}$. However, the exceptions can be handled using the same arguments that were presented in Section 3.4.1.

3.5 General large deviations inequality

Let $\{Z_n\}$ be a sequence of random vectors taking values in $\mathbb{R}^p$, and let μ_n denote the probability law of Z_n. Let Ψ_n be the cumulant-generating function of Z_n, and let $\{a_n\}$ be a sequence satisfying $a_n \to \infty$ as $n \to \infty$. Define $\Psi(\gamma)$ as in (3.10). The functions Ψ_n and Ψ are convex. Let $D_\Psi = \{\gamma \in \mathbb{R}^p : \Psi(\gamma) < \infty\}$ be the convex support set of Ψ

and note that $0 \in D_\Psi$.

Let I be the Fenchel-Legendre transform of Ψ as in (3.11). The classical Gärtner-Ellis theorem [Dembo and Zeitouni, 2009] establishes the inequality (3.9) for any closed measurable set E, under the condition that the origin belongs to the *interior* of D_Ψ. This condition will fail to hold in our setting, because we will consider situations in which D_Ψ is a subspace of $\mathbb{R}^p$. Thus, it is necessary to prove (3.9) under weaker conditions.

In the following, let $\mathcal{P}$ be the orthogonal projection operator onto the subspace D_Ψ, and define $\mathcal{P}E = \{\mathcal{P}u : u \in E\}$ to be the projection of any $E \subseteq \mathbb{R}^p$. Let $\mu_n^{\mathcal{P}}$ be the probability law of the random variable $\mathcal{P}Z_n$.

Our goal is to prove (3.9), for any closed measurable set E, under the assumption that $D_\Psi \neq \{0\}$ is a subspace of $\mathbb{R}^p$. This is accomplished in three steps with progressively fewer assumptions on E. In the first two steps (Lemma 3.12), the large deviations inequality is proved for $\mathcal{P}E$, with the first step making the additional assumption that this projected set is compact. The final step (Theorem 3.14) then proves the inequality for E.

Lemma 3.12. *Suppose that $D_\Psi \neq \{0\}$ is a subspace of $\mathbb{R}^p$, and $E \subseteq \mathbb{R}^p$ has the property that $\mathcal{P}E$ is compact and measurable. Then,*

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n^{\mathcal{P}}(\mathcal{P}E) \leq -\inf_{u\in\mathcal{P}E} I(u).$$

Proof: First notice that

$$I(u) = \sup_{\gamma\in\mathbb{R}^p} \{\langle\gamma, u\rangle - \Psi(\gamma)\} = \sup_{\gamma\in D_\Psi} \{\langle\gamma, u\rangle - \Psi(\gamma)\}$$

Let $I^\tau(u) = \min\left\{I(u) - \tau, \frac{1}{\tau}\right\}$ for $\tau > 0$. By definition of this function, for any $u \in \mathcal{P}E$ we can pick $\gamma^u \in D_\Psi$ for which $\langle \gamma^u, u\rangle - \Psi(\gamma^u) \geq I^\tau(\gamma^u)$. We can also pick ρ^u such that $\rho^u \|\gamma^u\| \geq \tau$ and let B_{u,ρ^u} be the closed ball of radius ρ^u centered at u.

By Chebyshev's inequality,

$$\mu_n^{\mathcal{P}}(G) = \mathbb{E}\left(1_{\{\mathcal{P}Z_n \in G\}}\right) \leq \mathbb{E}\left[\exp\left(\langle \gamma, \mathcal{P}Z_n\rangle - \inf_{u\in G}\langle \gamma, u\rangle\right)\right]$$

for any n, $\gamma \in \mathbb{R}^p$ and measurable $G \subseteq D_\Psi$. In particular,

$$\mu_n^{\mathcal{P}}(\mathcal{P}B_{u,\rho^u}) \leq \mathbb{E}\left[\exp\left(a_n\langle \gamma^u, \mathcal{P}Z_n\rangle\right)\right] \exp\left(-\inf_{u'\in\mathcal{P}B_{u,\rho^u}}\langle a_n\gamma^u, u'\rangle\right).$$

For any $u \in \mathcal{P}E$,

$$-\inf_{u'\in B_{u,\rho^u}}\langle a_n\gamma^u, u'\rangle \leq a_n\rho^u\|\gamma^u\| - a_n\langle\gamma^u, u\rangle \leq a_n\tau - a_n\langle\gamma^u, u\rangle,$$

whence

$$\begin{aligned}
\frac{1}{a_n}\log\mu_n^{\mathcal{P}}(\mathcal{P}B_{u,\rho^u}) &\leq \frac{1}{a_n}\log\mathbb{E}\left[\exp\left(a_n\langle\gamma^u, \mathcal{P}Z_n\rangle\right)\right] + \tau - \langle\gamma^u, u\rangle \\
&\leq \frac{1}{a_n}\log\mathbb{E}\left[\exp\left(\langle a_n\mathcal{P}\gamma^u, Z_n\rangle\right)\right] + \tau - \langle\gamma^u, u\rangle \qquad (3.34) \\
&= \frac{1}{a_n}\Psi_n\left(a_n\mathcal{P}\gamma^u\right) + \tau - \langle\gamma^u, u\rangle,
\end{aligned}$$

where (3.34) follows from the fact that $\mathcal{P}$ is self-adjoint.

Since $\mathcal{P}E$ is compact, we can select a finite covering from the open covering $\bigcup_{u\in\mathcal{P}E} B_{u,\rho^u}$ of $\mathcal{P}E$. Let N be the number of balls in this covering, and denote their

centers by u_i, $i = 1, ..., N$. For simplicity, let γ_i, ρ_i denote the corresponding γ^u, ρ^u values. Then,

$$\frac{1}{a_n} \log \mu_n^{\mathcal{P}} (\mathcal{P}E) \leq \frac{1}{a_n} \log N + \tau - \min_{1 \leq i \leq N} \left\{ \langle \gamma_i, u_i \rangle - \frac{1}{a_n} \Psi_n (a_n \mathcal{P} \gamma_i) \right\},$$

and we can take the limsup of both sides to obtain

$$\begin{aligned} \limsup_{n \to \infty} \frac{1}{a_n} \log \mu_n^{\mathcal{P}} (\mathcal{P}E) &\leq \tau - \min_{1 \leq i \leq n} \left\{ \langle \gamma_i, u_i \rangle - \limsup_{n \to \infty} \frac{1}{a_n} \Psi_n (a_n \mathcal{P} \gamma_i) \right\} \\ &= \tau - \min_{1 \leq i \leq n} \left\{ \langle \gamma_i, u_i \rangle - \Psi (\mathcal{P} \gamma_i) \right\}. \end{aligned}$$

Recalling the properties of γ_i, we arrive at

$$\limsup_{n \to \infty} \frac{1}{a_n} \log \mu_n^{\mathcal{P}} (\mathcal{P}E) \leq \tau - \min_{1 \leq i \leq n} I^{\tau} (\gamma_i) \leq \tau - \inf_{u \in \mathcal{P}E} I^{\tau} (u).$$

This holds for any $\tau > 0$, so we take $\tau \searrow 0$ to prove the desired result. □

Lemma 3.13. *Suppose that $D_\Psi \neq \{0\}$ is a subspace of $\mathbb{R}^p$, and $E \subseteq \mathbb{R}^p$ is closed and measurable. Then,*

$$\limsup_{n \to \infty} \frac{1}{a_n} \log \mu_n^{\mathcal{P}} (\mathcal{P}E) \leq - \inf_{u \in \mathcal{P}E} I (u).$$

Proof: Let $u_1, ..., u_\ell$ be a basis for the subspace D_Ψ, with $\ell < p$ being its dimensionality. Denote by μ_n^j the probability law of $\langle u_j, Z_n \rangle$.

Let $\gamma = a_n u_j$ and take some $\zeta > 0$. By Chebyshev's inequality,

$$\begin{aligned}\mu_n^j\left([\zeta,\infty)\right) &\leq \mathbb{E}\left[\exp\left(\langle a_n u_j, \mathcal{P}Z_n\rangle - \inf_{u:\langle u_j,u\rangle\geq\zeta}\langle a_n u_j, u\rangle\right)\right] \\ &\leq \mathbb{E}\left[\exp\left(a_n\langle u_j, \mathcal{P}Z_n\rangle\right)\right]\exp\left(-a_n\zeta\right),\end{aligned}$$

whence

$$\limsup_{n\to\infty}\frac{1}{a_n}\log\mu_n^j\left([\zeta,\infty)\right) \leq \Psi\left(u_j\right) - \zeta < \infty.$$

Consequently,

$$\lim_{\zeta\to\infty}\limsup_{n\to\infty}\frac{1}{a_n}\log\mu_n^j\left([\zeta,\infty)\right) = -\infty$$

for all $j = 1, ..., \ell$. Using symmetric arguments, one can also obtain

$$\lim_{\zeta\to\infty}\limsup_{n\to\infty}\frac{1}{a_n}\log\mu_n^j\left((-\infty,-\zeta]\right) = -\infty.$$

Now define the compact set $G_\zeta = \{u \in D_\psi : \langle u_j, u\rangle \in [-\zeta,\zeta]\ \forall j = 1, ..., \ell\}$. We then derive

$$\begin{aligned}&\lim_{\zeta\to\infty}\limsup_{n\to\infty}\frac{1}{a_n}\log\mu_n^{\mathcal{P}}\left(D_\psi \setminus G_\zeta\right) \\ \leq\ &\lim_{\zeta\to\infty}\limsup_{n\to\infty}\frac{1}{a_n}\log\sum_{j=1}^{\ell}\mu_n^j\left((-\infty,-\zeta]\right) + \mu_n^j\left([\zeta,\infty)\right) \\ \leq\ &\lim_{\zeta\to\infty}\limsup_{n\to\infty}\frac{1}{a_n}\log\left(2\ell\max_j\left\{\mu_n^j\left((-\infty,-\zeta]\right),\mu_n^j\left([\zeta,\infty)\right)\right\}\right) \\ =\ &-\infty, \qquad (3.35)\end{aligned}$$

where the first inequality uses a union bound together with the monotonicity of probability

measures.

Observing that $\mathcal{P}E \cap G_\zeta$ is compact, we can apply Lemma 3.12 to obtain

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n^{\mathcal{P}} \left(\mathcal{P}E \cap G_\zeta\right) \leq - \inf_{u \in \mathcal{P}E \cap G_\zeta} I\left(u\right) \leq - \inf_{u \in \mathcal{P}E} I\left(u\right).$$

On the other hand, $\mathcal{P}E \cap G_\zeta^c \subseteq D_\psi \setminus G_\zeta$, so

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n^{\mathcal{P}} \left(\mathcal{P}E \cap G_\zeta^c\right) \leq \limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n^{\mathcal{P}} \left(D_\psi \setminus G_\zeta\right).$$

Combining both inequalities, we find that

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n^{\mathcal{P}} \left(\mathcal{P}E\right) \leq 2 \max \left\{ - \inf_{u \in \mathcal{P}E} I\left(u\right), \limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n^{\mathcal{P}} \left(D_\psi \setminus G_\zeta\right) \right\}.$$

Taking $\zeta \to \infty$ and applying (3.35) yields the desired result. □

Theorem 3.14. *Suppose that $D_\Psi \neq \{0\}$ is a subspace of $\mathbb{R}^p$, and $E \subseteq \mathbb{R}^p$ is closed and measurable. Then,*

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n \left(E\right) \leq - \inf_{u \in E} I\left(u\right).$$

Proof: We rewrite (3.11) as

$$I\left(u\right) = \sup_{\gamma \in D_\Psi} \langle \gamma, u \rangle - \Psi\left(\gamma\right),$$

because $\Psi(\gamma)$ takes finite values only for $\gamma \in D_\psi$. Observe, however, that

$$\begin{aligned}\sup_{\gamma \in D_\Psi} \langle \gamma, u \rangle - \Psi(\gamma) &= \sup_{\gamma \in \mathbb{R}^p} \langle \mathcal{P}\gamma, u \rangle - \Psi(\gamma) \\ &= \sup_{\gamma \in \mathbb{R}^p} \langle \gamma, \mathcal{P}u \rangle - \Psi(\gamma) \\ &= I(\mathcal{P}u)\end{aligned}$$

because $\mathcal{P}$ is self-adjoint. Therefore, by Lemma 3.13,

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(E) \leq \limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n^{\mathcal{P}}(\mathcal{P}E) \leq -\inf_{u \in \mathcal{P}E} I(u) \leq -\inf_{u \in \mathcal{E}} I(u),$$

which completes the proof. □

We remark that the large deviations inequality can be recovered under the weaker condition $0 \notin \text{relint}(D_\psi)$, without requiring D_ψ to be a subspace of $\mathbb{R}^p$. However, this is beyond the needs of the present work and so we do not give the proof here.

3.6 Conclusion

We have presented a theoretical framework that leverages the connections between Gaussian process regression and approximation theory to derive new moderate deviations inequalities for different types of error probabilities. The utility of these results is demonstrated through two applications of broad interest: probabilities of pairwise errors between fixed errors of points, and tail probabilities for the estimation error of the minimum value. Furthermore, our results illustrate the effect of the kernel on the convergence rate.

It is difficult to say whether it is possible to improve on these bounds; perhaps this

also depends on the class of kernels that is chosen. The main limitation of this work is that, for purposes of tractability, we bound difficult posterior covariances by the much more tractable mesh norm. The mesh norm only measures the extent to which the design points are evenly spread out, and thus has limited ability to distinguish between different strategies for choosing the design points. We leave this problem for future work, noting that the results presented here are the first of their kind.

Chapter 4: Efficient Top-r Simultaneous Asymmetric Orthogonal Tensor Decomposition

4.1 Introduction

Tensor decomposition has been a key spectral algorithm for solving many ML problems. In recent years, rapid improvements in hardware have driven new emerging applications in biomedicine, signal processing, data mining and computer vision. Specifically, it works for latent variable models, a very broad class of probabilistic models encompassing Gaussian mixture models, hidden Markov models, and latent Dirichlet allocation, all widely used in machine learning. Tensor decomposition can be used to develop estimators for the high dimensional structure that is complex and hidden. In solving such models, the method of moments Anandkumar et al. [2014a], Hall [2005] relates the observed data moments with model parameters using tensor CANDECOMP/PARAFAC (CP) decompositions Kolda and Bader [2009]. Consistent model parameter estimators are obtained through orthogonal tensor decompositions Anandkumar et al. [2014a].

A few challenges remain unsettled in this context. **First**, these consistent estimators through orthogonal tensor decompositions have to be efficiently computed via orthogonal decomposition of a tensor of observed moments. This efficiency is especially valuable

in high-dimensional problems where the number of cross-feature moments can be large. **Second**, a full recovery is not always attractive - we may only care about the dominant/top components in the hidden structure. **Moreover**, in these applications, we may observe an empirical moment $\widehat{\mathcal{T}}$, a noisy version of the data moment $\mathcal{T}$. It is assumed that $\widehat{\mathcal{T}}$ can be decomposed as $\widehat{\mathcal{T}} = \mathcal{T} + \mathbf{\Phi}$, where $\mathbf{\Phi}$ is the noise tensor. Therefore, the core objective is to find robust-to-noise methods that provide guaranteed recovery of top components of $\mathcal{T}$ using $\widehat{\mathcal{T}}$, within a small number of iterations.

Consider a 3-order underlying tensor $\mathcal{T}$ with components $\mathbf{A}$,$\mathbf{B}$, $\mathbf{C}$, and let $\mathcal{T} = \sum_{i=1}^{R} \lambda_i \mathbf{a}_i \otimes \mathbf{b}_i \otimes \mathbf{c}_i$, where $\mathbf{a}_i$, $\mathbf{b}_i$, $\mathbf{c}_i$ are the columns of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ respectively. If $\mathcal{T}$ is *symmetric*, it permits a symmetric CP decomposition $\mathbf{A} = \mathbf{B} = \mathbf{C}$. If $\mathcal{T}$ is asymmetric, $\mathcal{T}$ must be decomposed via an *asymmetric* decomposition $\mathbf{A} \neq \mathbf{B} \neq \mathbf{C}$.

Efficient Convergence Rate Kolda [2015a] argues that symmetric orthogonal tensor decomposition is trivial. A method is proposed in this paper to compute an orthogonal decomposition of an m-way d-dimensional symmetric tensor. The problem is then reduced to an $d \times d$ symmetric matrix eigen-problem. The convergence rate of the matrix eigen-problem solver is inevitably linear ($O(\log \frac{1}{\epsilon})$) for an ϵ-close recovery, slower than the convergence rate of tensor power methods (convergence rate increases with tensor order m, for instance, a 3-way tensor achieves quadratic convergence rate $O(\log \log \frac{1}{\epsilon})$).

Simultaneous Recovery Popular tensor decomposition methods such as tensor power method, although achieve quadratic convergence rate for 3-way tensors, recovers components one by one. Unlike previous schemes based on deflation methods that recover

factors sequentially Anandkumar et al. [2014b], our scheme recovers any number of top components simultaneously even when R is unknown. This is a more practical setting. In numerous machine learning settings, data is generated in real-time, and sequential recovery of factors may be inapplicable under such online settings. Prior work Wang and Lu [2017] considers a simultaneous subspace iteration, but is only limited to symmetric tensors.

Asymmetric Tensors Many alternative methods exist that are popular in the symmetric case (e.g. Brachat et al. [2010], Kolda [2015b], Nie [2017]), but the symmetric assumption required by these methods is restrictive. In most applications, multi-view models or HMMs, in which information is asymmetric along different modes, are needed. Decomposition of symmetric tensors is easier than that of asymmetric ones, as the constraints of symmetric entries vastly reduce the number of parameters in the CP decomposition problem. There are prior work Anandkumar et al. [2014a], Anandkumar et al. [2016], Goyal et al. [2014], Sharan and Valiant [2017], Wang and Lu [2017] on decomposing symmetric tensors with identical components across modes. All these methods require multiple random sampling initializations which inevitably induce convergence of the algorithms, only with high probability.

In this paper, we consider simultaneous top r components recovery of asymmetric tensors with unknown R number of orthonormal components. Our goal is to recover top r components simultaneously and almost surely when noiseless. Our *Slicing Initialized Alternating Subspace Iteration* (s-ASI) uses a tensor subspace iteration method, i.e., orthogonalized alternating least square (o-ALS).

4.1.1 Summary of contribution

Contribution to Asymmetric Tensor Decomposition We provide the first guaranteed decomposition algorithm, Slicing Initialized Alternating Subspace Iteration (s-ASI), for asymmetric tensors with a convergence rate $O(\log\log\frac{1}{\epsilon})$ independent of the rank and dimension. Thanks to Slice-Based Initialization using only $O(1/\log(\frac{\lambda_r}{\lambda_{r+1}}))$ steps, s-ASI recovers the top r components (corresponding to the largest r singular values) simultaneously with probability 1 under the noiseless case when R is unknown. Our s-ASI is also robust to noise smaller than $\min\{\frac{\sqrt{2}}{8}\frac{\Delta\epsilon}{\sqrt{R}}, \delta_0\frac{\lambda_r^2-\lambda_{r+1}^2}{8\|\lambda\|}, \delta_0\frac{\Delta}{2\sqrt{d}}\}$, where $\Delta = \min_r(\lambda_r - \lambda_{r+1})$ denotes the spectral gap of the tensor, d the dimension and δ_0 a constant proportional to the failure probability of initialization.

Contribution to Symmetric Tensor Decomposition Our Slice-Based Initialization procedure applies to symmetric orthogonal tensor decomposition to (1) provide an initialization that guarantees convergence to top r components almost surely when the tensor is noiseless (in contrast to the random sampling based initialization method Wang and Lu [2017] which leads to convergence with some high probability); (2) improve the robustness of the algorithm by allowing larger noise $\min\{O(\frac{\Delta\epsilon}{\sqrt{R}}), O(\delta_0\frac{\Delta}{\sqrt{d}})\}$, in contrast to the state-of-the-art noise level $\min\{O(\frac{\Delta\epsilon}{\sqrt{R}}), O(\delta_0\frac{\Delta^2}{\sqrt{dR}}))\}$ allowed. Here we use the fact that the bound can be loosened by replacing $\lambda_r^2 - \lambda_{r+1}^2$ with Δ^2.

Theorem 4.1 (Informal s-ASI Convergence Guarantee)**.** *Let a tensor permit a noisy orthogonal CP decomposition form* $\widehat{\mathcal{T}} = \sum_{i=1}^{R}\lambda_i\mathbf{a}_i\otimes\mathbf{b}_i\otimes\mathbf{c}_i + \mathbf{\Phi}$ *with bounded noise, where* λ_i *are in descending order. After running* $O(\log(\log\frac{1}{\epsilon}))$ *steps of tensor subspace iteration in our Alternating Subspace Iteration (Procedure 1), the estimated* i^{th} *component* $\mathbf{a}_i^*$

converges to the i-th component $\mathbf{a}_i$ with high probability $\|\mathbf{a}_i - \mathbf{a}_i^*\| \leq \epsilon$ for $\forall 1 \leq i \leq r$.

Note that the results are identifiable up to sign flips only. In contrast to rank-1 methods, which are identifiable up to both sign flip and column permutation, our s-ASI identifies the top-r components with largest λ_i in the correct order. We shall also emphasize that $1 \leq r \leq d$ can be an arbitrary number without any required knowledge of R. This fact is critical because it saves computing resources from recovering insignificant or unwanted components.

4.2 Related works

Rank-1 methods Both popular rank-1 power methods Anandkumar et al. [2014a], Wang and Anandkumar [2016] (on orthogonal symmetric tensors using random initialization and deflation) and rank-1 ALS Anandkumar et al. [2014b] (on incoherent tensors via optimizing individual mode of the factors while fixing all other modes, and alternating between the modes) require recovery of **all** R components sequentially to determine the top r components. This is becuase top components are not necessarily first recovered. Therefore the convergence rates will inevitably contain a factor of R making their method slower than our s-ASI.

Rank-r methods (1) **Comparison with rank-r power method**. Wang et al. Wang and Lu [2017] use subspace iteration and prove the simultaneous convergence of the top-k singular vectors for orthogonal symmetric tensors. A sampling-based procedure is used for initialization. Their sampling-based initialization inevitably introduces a high probability bound even when the observed data is noiseless. (2) **Comparison with rank-r**

orthogonal ALS. Convergence of a variant of ALS using QR decomposition Sharan and Valiant [2017] with random initialization for symmetric tensors has been proven to require number of iterations with a factor of R. Their method converges to the top r components only when the rank R is known and $r = R$. Their convergence bound of sequential analysis is found to be loose.

Gradient-based methods Stochastic gradient descent is used to handle tensor decomposition problems. In Ge et al. [2015], an objective function for tensor decomposition is proposed where all the local optima are globally optimal. However, the polynomial convergence rate is slower than the double exponential rate achieved in our paper.

Matrix-based methods Tomasi and Bro [2006] provides a general survey on some early efforts on matrix-based methods. Most are based on reduction to matrix decomposition (including subroutines that solves CP decomposition for two-slice tensors through joint diagonalization(Domanov and Lathauwer [2014] Roemer and Haardt [2008])). Our method improves upon the line of work mentioned due to the following reasons. (a) We propose a noise-robust algorithm that fast converges to top-r components. In contrast, neither Domanov and Lathauwer [2014] nor Roemer and Haardt [2008] present a convergence rate analysis or robustness analysis under noise. (b) Tomasi and Bro [2006] also discussed several types of trilinear decomposition methods (TLD), which call matrix decompositive subroutines that limit their convergence rates to be slower than ours. For others mentioned in Tomasi and Bro [2006], our method outperforms them in terms of either convergence rate, memory expense, resistance of over-factoring, or ability of simultaneous recovery of top-r components.

It is empirically shown in Faber et al. [2003] that a preliminary version of ALS

outperforms a series of trilinear decomposition methods (DTLD, ATLD, SWATLD). Our algorithm outperforms the state-of-the-art ALS method in experiments. The method in De Lathauwer et al. [2004] also fulfills simultaneous recovery, but it involves interations essentially using eigenvalue decomposition and a step of minimizing a cost function, for which the convergence is not ensured to be global. More recent works in this direction include Kuleshov et al. [2015] and Pimentel-Alarcón [2016]. Kuleshov et al Kuleshov et al. [2015] proposed a sophisticated way of projection such that the gaps of eigenvalues are preserved with high probability. However there is no guarantee of top r recovery.

Matrix-decomposition-based methods generally have a linear (logarithmic) convergence rate. **Eigen-decomposition based methods** are promising, for example the one introduced in Kolda [2015a]. Even so, our method still wins its place by requiring less iterations ($O(\log(\log \frac{1}{\epsilon}))$ versus $O(\log \frac{1}{\epsilon})$ from Kolda [2015a]), and possibly even for less computational complexity. The total computation cost of our method is $O(d^3 r \log(\log \frac{1}{\epsilon}))$ whereas the method in Kolda [2015a] takes $O(d^3 \log \frac{1}{\epsilon})$, dominated by the cost for full eigen-decomposition. Our cost would indeed be much less when the number of components wanted is a constant $r = O(1)$ in terms of dimension d and recovery precision ϵ. Very importantly, we remark that doing a truncated eigen-decomposition will not necessarily recover top components of the tensor. Another advantage of our method is the analysis for noise tolerance for (a)symmetric tensors, which is either not allowed or missing in the eigen-decomposition based methods.

4.3 Tensor & subspace iteration preliminaries

Let $[n] := \{1, 2, \ldots, n\}$. For a vector $\mathbf{v}$, denote the i^{th} element as v_i. For a matrix $\mathbf{M}$, denote the i^{th} row as $\mathbf{m}^i$, j^{th} column as $\mathbf{m}_j$, and $(i, j)^{\text{th}}$ element as m_{ij}. Denote the first r columns of matrix $\mathbf{M}$ as $\mathbf{M}_r$. An n-order (number of dimensions, a.k.a. modes) tensor, denoted as $\mathcal{T}$, is a multi-dimensional array with n dimensions. For a 3-order tensor $\mathcal{T}$, its $(i, j, k)^{\text{th}}$ entry is denoted by T_{ijk}. A tensor is called *cubical* if every mode is of the same size. A cubical tensor is called *supersymmetric* (or simply refered as symmetric thereafter) if its elements remain constant under any permutation of the indices.

Tensor product is also known as outer product. For $\mathbf{a} \in \mathbb{R}^m, \mathbf{b} \in \mathbb{R}^n$ and $\mathbf{c} \in \mathbb{R}^p$, $\mathbf{a} \otimes \mathbf{b} \otimes \mathbf{c}$ is a $m \times n \times p$ sized 3-way tensor with $(i, j, k)^{\text{th}}$ entry being $a_i b_j c_k, \forall 1 \leq i \leq m, 1 \leq j \leq n, 1 \leq k \leq p$.

Multilinear Operation The tensor-vector/matrix multilinear operation of $\mathcal{T}$ and matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ is defined as: $[\mathcal{T}(\mathbf{A}, \mathbf{B}, \mathbf{C})]_{ijk} = \sum_{a,b,c} \mathcal{T}_{abc} \mathbf{A}_{ai} \mathbf{B}_{bj} \mathbf{C}_{ck}$. The tensor-vector multiplication is defined similarly.

Tensor operator norm The operator norm for tensor $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is defined as $\|\mathcal{T}\|_{\text{op}} = \max_{\mu_i \in \mathbb{R}^{d_i} \setminus \{\mathbf{0}\}, i=1,2,3} \frac{|\mathcal{T}(\mu_1, \mu_2, \mu_3)|}{\|\mu_1\| \cdot \|\mu_2\| \cdot \|\mu_3\|}$.

Matricization is the process of reordering the elements of an N-way tensor into a matrix. The mode-n matricization of a tensor $\mathcal{T} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ is denoted by $\mathcal{T}_{(n)}$ and arranges the mode-n fibers Kolda and Bader [2009] to be the columns of the resulting matrix, i.e., the $(i_1, i_2, ..., i_N)^{\text{th}}$ element of the tensor maps to the $(i_n, j)^{\text{th}}$ element of the matrix, where $j = 1 + \sum_{k=1, k \neq n}^{N} (i_k - 1) \prod_{m=1, m \neq n}^{k-1} I_m$.

Khatri-rao product $\mathbf{A} \odot \mathbf{B} := \begin{bmatrix} a_{11}\mathbf{b}_1 \cdots a_{1p}\mathbf{b}_p \\ \vdots \quad \ddots \quad \vdots \\ a_{m1}\mathbf{b}_1 \cdots a_{mp}\mathbf{b}_p \end{bmatrix}$, for $\mathbf{A} \in \mathbb{R}^{m \times p}$, $\mathbf{B} \in \mathbb{R}^{n \times p}$.

Tensor CP decomposition A tensor $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ has *CP decomposition* if the tensor could be expressed exactly as a sum of R rank-one components, i.e. $\exists\ \mathbf{\Lambda}$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ such that $\mathcal{T} = \sum_{i=1}^{R} \lambda_i \mathbf{a}_i \otimes \mathbf{b}_i \otimes \mathbf{c}_i$, where R is a positive integer, $\mathbf{\Lambda} = \mathsf{Diag}([\lambda_1, \lambda_2, \cdots, \lambda_R])$, $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_R] \in \mathbb{R}^{d_1 \times R}$, $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_R] \in \mathbb{R}^{d_2 \times R}$ and $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_R] \in \mathbb{R}^{d_3 \times R}$. If so, we donote the CP decomposition as $\mathcal{T} = [\![\mathbf{\Lambda}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!]$ and call $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ factors of this CP decomposition. The *rank* of $\mathcal{T}$ is the smallest number of rank-one components that sum to $\mathcal{T}$.

Subspace similarity The definition follows Zhu and Knyazev [2013].

Definition 4.1 (Subspace Similarity)**.** *Let S_1, S_2 be two m-dimension proper subspaces in $\mathbb{R}^n$ spanned respectively by columns of two basis matrices $\mathbf{M}_1, \mathbf{M}_2$. Let $\mathbf{M}_2^c$ be the basis matrix for the complement subspace of S_2. The principal angle θ formed by S_1 and S_2 is* $\cos(\theta) = \min_{\mathbf{y} \in \mathbb{R}^m} \frac{\|\mathbf{M}_1^\top \mathbf{M}_2 \mathbf{y}\|}{\|\mathbf{M}_2 \mathbf{y}\|} = \sigma_{min}(\mathbf{M}_1^\top \mathbf{M}_2)$, $\sin(\theta) = \max_{\mathbf{y} \in \mathbb{R}^{n-m}} \frac{\|\mathbf{M}_1^\top \mathbf{M}_2^c \mathbf{y}\|}{\|\mathbf{M}_2^c \mathbf{y}\|} = \sigma_{max}(\mathbf{M}_1^\top \mathbf{M}_2^c)$, $\tan(\theta) = \frac{\sin(\theta)}{\cos(\theta)} = \frac{\sigma_{max}(\mathbf{M}_1^\top \mathbf{M}_2^c)}{\sigma_{min}(\mathbf{M}_1^\top \mathbf{M}_2)}$, *where $\sigma_{min}(\cdot)$ / $\sigma_{max}(\cdot)$ denotes the smallest / greatest singular value of a matrix.*

4.4 Asymmetric tensor decomposition model

Consider a rank-R asymmetric tensor $\mathcal{T} \in \mathbb{R}^{d \times d \times d}$ with latent factors $\mathbf{\Lambda}$, $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$

$$\mathcal{T} = [\![\mathbf{\Lambda}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \equiv \sum_{i=1}^{R} \lambda_i \mathbf{a}_i \otimes \mathbf{b}_i \otimes \mathbf{c}_i \tag{4.1}$$

where $\mathbf{\Lambda}$=Diag($[\lambda_1, \cdots, \lambda_R]$), $\mathbf{A}$=$[\mathbf{a}_1, \ldots, \mathbf{a}_R] \in \mathbb{R}^{d \times R}$ and $\mathbf{A}^\top\mathbf{A}$=$\mathbf{I}$ (similarly for $\mathbf{B}$, $\mathbf{C}$). Without loss of generality, we assume $\lambda_1 > \lambda_2 > \cdots > \lambda_R > 0$. Our analysis applies to general order-n symmetric and asymmetric tensors. In this paper, $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are all orthonormal matrices, and therefore the tensor we find the CP decomposition on has a unique orthogonal decomposition, based on Kruskal's condition Kruskal [1977].

Our goal is to discover a CP decomposition with R orthogonal components that best approximates the observed $\widehat{\mathcal{T}}$. This can be formulated as solving the following optimization problem:

$$\arg \min_{\mathbf{\Lambda}^*, \mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*} \left\| \widehat{\mathcal{T}} - [\![\mathbf{\Lambda}^*; \mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*]\!] \right\|_{\mathsf{F}}^2$$
$$\text{s.t. } \Lambda^*_{i,j} = 0, \forall i \neq j, \mathbf{A}^{*\top}\mathbf{A}^* = \mathbf{I}, \mathbf{B}^{*\top}\mathbf{B}^* = \mathbf{I}, \mathbf{C}^{*\top}\mathbf{C}^* = \mathbf{I}$$

We denote the estimated singular values and factor matrices as $\mathbf{\Lambda}^*$, $\mathbf{A}^*$, $\mathbf{B}^*$ and $\mathbf{C}^*$ respectively.

4.4.1 Difficulty of asymmetric tensor decomposition

Asymmetric tensor decomposition is more difficult than symmetric tensor decomposition due to the following reasons: (1) the number of parameters required to be estimated is a factor of the tensor order more than the symmetric tensor decomposition (2) the missing symmetry imposes additional difficulty for simultaneous recovery of top-r components of the tensor.

Symmetrization Instability Existing works often assume that an asymmetric tensor can be symmetrized by a multilinear operation, i.e., $\mathcal{T}(\mathbf{M}_a, \mathbf{M}_b, \mathbf{I})$ becomes symmetric, and thus only prove convergence of symmetric tensor decomposition. Here the symmetrization matrices $\mathbf{M}_a = \mathcal{T}(\mathbf{b}, \mathbf{I}, \mathbf{I})^\top \mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{a})^{-1}$ and $\mathbf{M}_b = \mathcal{T}(\mathbf{I}, \mathbf{b}, \mathbf{I})^\top (\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{a})^\top)^{-1}$ with $\mathbf{a}$ and $\mathbf{b}$ sampled from a unit sphere. For a proof of the symmetrization, see Appendix 4.7.2. However, the computation of $\mathbf{M}_a$ and $\mathbf{M}_b$ can be unstable due to the inversion of $\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{a})^{-1}$. Specifically, the inversion of $\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{a})$ can be ill-conditioned, i.e., the condition number $\kappa(\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{a})) = \frac{\max_i \lambda_i(\mathbf{a}^\top \mathbf{c}_i)}{\min_i \lambda_i(\mathbf{a}^\top \mathbf{c}_i)}$ can be high. Therefore, we consider the direct asymmetric tensor decomposition.

4.5 Simultaneous asymmetric tensor decomposition

One way to solve the trilinear optimization problem in Equation (**??**) is through the alternating least square (ALS) method Carroll and Chang [1970], Harshman [1970], Kolda and Bader [2009]. The ALS (without orthognalization) approach fixes $\mathbf{B}, \mathbf{C}$ to compute a closed form solution for $\mathbf{A}$, then fixes $\mathbf{A}, \mathbf{C}$ for $\mathbf{B}$, and fixes $\mathbf{A}, \mathbf{B}$ for $\mathbf{C}$. The alternating updates are repeated until the convergence criterions are satisfied. By fixing all but one factor matrix, the problem reduces to a linear least-squares problem over the matricized tensor

$$\arg\min_{\mathbf{A}^*, \mathbf{\Lambda}^*} \|\widehat{\mathcal{T}}_{(1)} - \mathbf{A}^* \mathbf{\Lambda}^* (\mathbf{C}^* \odot \mathbf{B}^*)^\top\|_{\mathsf{F}}^2, \tag{4.2}$$

where there exists a closed form solution $\mathbf{A}^* \mathbf{\Lambda}^* = \widehat{\mathcal{T}}_{(1)}[(\mathbf{C}^* \odot \mathbf{B}^*)^\top]^\dagger$, using the pseudo-inverse. ALS converges quickly and is usually robust to noise in practice. However, the convergence theory of ALS for asymmetric tensors is not well understood. We fill

the gap in this paper by introducing an *alternating subspace iteration* (ASI) as shown in Algorithm 1, for asymmetric tensors.

We provide the convergence rate proof of our s-ASI for asymmetric tensor using two steps. **(1)** Under some r-*sufficient initialization condition* (defined in Definition 4.2), we prove an $O(\log(\log(\frac{1}{\epsilon})))$ convergence rate of ASI (Algorithm 1). **(2)** We propose a *Slice-Based Initialization* (Algorithm 2), and prove that after $O(1/\log\frac{\lambda_r}{\lambda_{r+1}})$ steps of matrix subspace iteration, r-sufficient initialization condition is satisfied. We call our algorithm *Slicing Initialized Alternating Subspace Iteration* (s-ASI).

4.5.1 ASI under r-sufficient initialization condition

We define the sufficient initialization condition in Definition 4.2, under which our Alternating Subspace Iteration algorithm is guaranteed to converge to the true factors of the tensor $\mathcal{T}$.

Definition 4.2 (r-Sufficient Initialization Condition)**.** *The r-sufficient initialization condition is satisfied if* $\tan\left(\mathbf{A}_r, \mathbf{Q}_{\mathbf{A}_r}^{(0)}\right) < 1$, $\tan\left(\mathbf{B}_r, \mathbf{Q}_{\mathbf{B}_r}^{(0)}\right) < 1$, *and* $\tan\left(\mathbf{C}_r,\ \mathbf{Q}_{\mathbf{C}_r}^{(0)}\right) < 1$.

Under a satisfaction of the r-sufficient initialization condition in Definition 4.2, we update the components $\mathbf{Q}_A^{(k+1)}$, $\mathbf{Q}_B^{(k+1)}$ and $\mathbf{Q}_C^{(k+1)}$ as in line 3,4,5 of Algorithm 1. We save on expensive matrix inversions over $(\mathbf{Q}_C^{(k)} \odot \mathbf{Q}_B^{(k)})$ as $(\mathbf{Q}_C^{(k)} \odot \mathbf{Q}_B^{(k)}) = [(\mathbf{Q}_C^{(k)} \odot \mathbf{Q}_B^{(k)})^\top]^\dagger$ due to the orthogonality of $\mathbf{Q}_B^{(k)}$ and $\mathbf{Q}_C^{(k)}$. We obtain the following conditional convergence theorem.

Theorem 4.2 (Noiseless Conditional Simultaneous Convergence)**.** *Under the r-sufficient initialization condition in definition 4.2 and noiseless scenario, after $K = O(\log(\log(\frac{1}{\epsilon})))$*

Algorithm 1: Alternating Subspace Iteration (Alternating Subspace Iterationshort) for Asymmetric Tensor Decomposition

Input: $d \times d \times d$ sized tensor $\widehat{\mathcal{T}}$, a tentative rank r, precision ϵ
Output: $\mathbf{\Lambda}^*, \mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*$, such that $\|\mathbf{A}_r - \mathbf{A}^*\|, \|\mathbf{B}_r - \mathbf{B}^*\|, \|\mathbf{C}_r - \mathbf{C}^*\| \leq \epsilon$

Initialize $\mathbf{Q}_A^{(0)}, \mathbf{Q}_B^{(0)}, \mathbf{Q}_C^{(0)}$ through Algorithm 2
for $k = 0$ ***to*** $K = O(\log(\log \frac{1}{\epsilon}))$ **do**
  $\mathbf{Q}_A^{(k+1)}\mathbf{R}_A^{(k+1)} \leftarrow \mathsf{QR}\left(\widehat{\mathcal{T}}_{(1)}(\mathbf{Q}_C^{(k)} \odot \mathbf{Q}_B^{(k)})\right)$
  $\mathbf{Q}_B^{(k+1)}\mathbf{R}_B^{(k+1)} \leftarrow \mathsf{QR}\left(\widehat{\mathcal{T}}_{(2)}(\mathbf{Q}_C^{(k)} \odot \mathbf{Q}_A^{(k+1)})\right)$
  $\mathbf{Q}_C^{(k+1)}\mathbf{R}_C^{(k+1)} \leftarrow \mathsf{QR}\left(\widehat{\mathcal{T}}_{(3)}(\mathbf{Q}_B^{(k+1)} \odot \mathbf{Q}_A^{(k+1)})\right)$
end
$(\mathbf{\Lambda}^*, \mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*) \leftarrow$ Algorithm 4($\widehat{\mathcal{T}}$, r, $\mathbf{Q}_A^{(K)}$, $\mathbf{Q}_B^{(K)}$, $\mathbf{Q}_C^{(K)}$)

steps, our Alternating Subspace Iteration in Algorithm 1 recovers the estimates of the factors $\mathbf{a}_i^*$, $\mathbf{b}_i^*$, *and* $\mathbf{c}_i^*$ *that correspond to the top-*r *true components with largest* λ_i *up to sign flip, i.e.,* $\|\mathbf{a}_i - \mathbf{a}_i^*\|^2 \leq 2\epsilon$, $\forall 1 \leq i \leq r$. *Similarly for* $\mathbf{b}_i^*$, $\mathbf{c}_i^*$, $\forall 1 \leq i \leq r$.

Theorem 4.2 guarantees that the estimated factors recovered using Alternating Subspace Iterationshort converges to the true factors $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ when noiseless. We also provided the guarantee for the noisy case in Section 4.7. The convergence rate of Alternating Subspace Iteration is $\log(\log(\frac{1}{\epsilon}))$ when the r-sufficient initialization condition is satisfied. The convergence result requires careful manipulation of three different modes. Most ALS methods assume a relaxation to asymmetric tensors, however the existing works only provide convergence results for symmetric tensors. Our work closes the gap between theory and practice. The proof sketch is in Appendix 4.7.3. We now propose a novel initialization method in Algorithm 2 which guarantees that the r-Sufficient Initialization Condition is satisfied.

Algorithm 2: Slice-Based Initialization

Input: Tensor $\widehat{\mathcal{T}}$, r
Output: $\mathbf{Q}_A^{(0)}, \mathbf{Q}_B^{(0)}, \mathbf{Q}_C^{(0)}$

$\mathbf{e}_i \leftarrow i^{\text{th}}$ column of identity matrix
if $\widehat{\mathcal{T}}$ *is asymmetric* **then**
  $\mathbf{M}^A \leftarrow \sum_{i=1}^{d} \widehat{\mathcal{T}}(\mathbf{I}, \mathbf{I}, \mathbf{e}_i)\widehat{\mathcal{T}}(\mathbf{I}, \mathbf{I}, \mathbf{e}_i)^\top$
  $\mathbf{M}^B \leftarrow \sum_{i=1}^{d} \widehat{\mathcal{T}}(\mathbf{e}_i, \mathbf{I}, \mathbf{I})\widehat{\mathcal{T}}(\mathbf{e}_i, \mathbf{I}, \mathbf{I})^\top$
  $\mathbf{M}^C \leftarrow \sum_{i=1}^{d} \widehat{\mathcal{T}}(\mathbf{I}, \mathbf{e}_i, \mathbf{I})^\top\widehat{\mathcal{T}}(\mathbf{I}, \mathbf{e}_i, \mathbf{I})$
else
  $\mathbf{M}^A \leftarrow \widehat{\mathcal{T}}(\mathbf{I}, \mathbf{I}, \mathbf{v}^C)$ // $v_i^C = \texttt{trace}(\widehat{\mathcal{T}}(\mathbf{I}, \mathbf{I}, \mathbf{e}_i))$
  $\mathbf{M}^B \leftarrow \widehat{\mathcal{T}}(\mathbf{v}^A, \mathbf{I}, \mathbf{I})$ // $v_i^A = \texttt{trace}(\widehat{\mathcal{T}}(\mathbf{e}_i, \mathbf{I}, \mathbf{I}))$
  $\mathbf{M}^C \leftarrow \widehat{\mathcal{T}}(\mathbf{I}, \mathbf{v}^B, \mathbf{I})^\top$ // $v_i^B = \texttt{trace}(\widehat{\mathcal{T}}(\mathbf{I}, \mathbf{e}_i, \mathbf{I}))$
end
$\mathbf{Q}_A^{(0)} \leftarrow$ output of Algorithm 3 on $\mathbf{M}^A$
$\mathbf{Q}_B^{(0)} \leftarrow$ output of Algorithm 3 on $\mathbf{M}^B$
$\mathbf{Q}_C^{(0)} \leftarrow$ output of Algorithm 3 on $\mathbf{M}^C$

4.5.2 r-Sufficient Initialization: Slice-Based Initialization and Matrix Subspace Iteration

We provide a guaranteed r-Sufficient Initialization $\mathbf{Q}_A^{(0)}, \mathbf{Q}_B^{(0)}, \mathbf{Q}_C^{(0)}$ using a 2-step procedure:

- Prepare matrix $\mathbf{M}^A$ ($\mathbf{M}^B$, $\mathbf{M}^C$) such that the left eigenspace is the column space of $\mathbf{A}$ ($\mathbf{B}$, $\mathbf{C}$). Unlike in Sharan and Valiant [2017] or Wang and Lu [2017], Algorithm 2 recovers $\mathbf{M}^A$ with preserved order of tensor components.

- Recover r-sufficient $\mathbf{Q}_A^{(0)}$ (same for $\mathbf{Q}_B^{(0)}$ and $\mathbf{Q}_C^{(0)}$) from the matrices above, achieved by Algorithm 3 almost surely in the noiseless case (the discussion of noisy setting is deferred to section 4.7).

We assume a gap between the r^{th} and the $(r+1)^{\text{th}}$ singular values for all $r \leq$

R. Lemma 4.13 in Appendix 4.7.3.4 provides the key intuition behind our initialization procedure. Lemma 4.13 shows that given a matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$, matrix subspace iteration in Algorithm 3 recovers the left eigenspace spanned by eigenvectors of $\mathbf{M}$ corresponding to p largest eigenvalues. Therefore, matrix subspace iteration provides insight into how the factors should be initialized. It suggests that as long as we find a matrix whose left eigenspace is the column space of $\mathbf{A}$, we can use matrix subspace iteration to prepare an initialization for Alternating Subspace Iterationshort.

Algorithm 3: Matrix Subspace Iteration

Input: Matrix $\mathbf{M}$, r
Output: Left invariant subspace approximation $\mathbf{Q}^{(J)}$
Initialize random orthogonal $\mathbf{Q}^{(0)} \in \mathbb{R}^{d \times r}$ from *Haar* distribution Mezzadri [2006]

for $j = 1$ ***to*** $J = O(\log(C)/\log(|\frac{\lambda_r}{\lambda_{r+1}}|))$ **do**
  $\mathbf{Q}^{(j)}\mathbf{R}^{(j)} \leftarrow \mathsf{QR}\left(\mathbf{M}\mathbf{Q}^{(j-1)}\right)$
end

Theorem 4.3 (Noiseless)**.** *Assume that $C \geq 1$ (otherwise r-Sufficient Initialization Condition is met after one iteration), after we run Algorithm 2 and 3 with*

$$J = O(\log(C)/\log(|\frac{\lambda_r}{\lambda_{r+1}}|))$$

steps, we guarantee under noiseless scenario, up to sign flip only $\tan\left(\mathbf{A}_r, \mathbf{Q}_A^{(0)}\right) < 1$, *same for* $\mathbf{Q}_B^{(0)}$ *and* $\mathbf{Q}_C^{(0)}$.

Theorem 4.3 guarantees that r-Sufficient Initialization Condition (Definition 4.2) is satisfied after $O(\log(C)/\log(|\frac{\lambda_r}{\lambda_{r+1}}|))$ steps of matrix subspace iteration. The proof of Theorem 4.3 (appendix) follows directly from Lemma 4.13 by setting the convergence

tolerance to 1.

Algorithm 4: Singular Value Computation

Input: $\widehat{\mathcal{T}}$, r, $\mathbf{Q}_A^{(K)}$, $\mathbf{Q}_B^{(K)}$, $\mathbf{Q}_C^{(K)}$
Output: $\mathbf{\Lambda}^*, \mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*$

for $i = 1$ ***to*** r **do**
  $\mathbf{a}_i^*, \mathbf{b}_i^*, \mathbf{c}_i^* \leftarrow$ the i^{th} column of $\mathbf{Q}_A^{(K)}$, $\mathbf{Q}_B^{(K)}$, $\mathbf{Q}_C^{(K)}$ respectively
  $\lambda_i^* \leftarrow \widehat{\mathcal{T}}(\mathbf{a}_i^*, \mathbf{b}_i^*, \mathbf{c}_i^*)$
end
$\mathbf{\Lambda}^* \leftarrow \mathsf{Diag}(\lambda_1^*, \cdots, \lambda_r^*)$, $\mathbf{A}^* \leftarrow \mathbf{Q}_A{}^{(K)}$, $\mathbf{B}^* \leftarrow \mathbf{Q}_B^{(K)}$, $\mathbf{C}^* \leftarrow \mathbf{Q}_C^{(K)}$

4.6 Slice-Based Initialization

For matrix subspace iteration in Algorithm 3 to work, we prepare a matrix that spans the space of eigenvectors of $\mathbf{A}$ using *Slice-Based Initialization* in Algorithm 2 for symmetric and asymmetric tensors. matrix subspace iteration is on $\widehat{\mathcal{T}}(\mathbf{I}, \mathbf{I}, \mathbf{v}^C)$ where $\mathbf{v}_i^C = \text{trace}(\widehat{\mathcal{T}}(\mathbf{I}, \mathbf{I}, \mathbf{e}_i)), \forall i \in [d]$ for symmetric tensor, and is on $\sum_{i=1}^{d} \widehat{\mathcal{T}}(\mathbf{I}, \mathbf{I}, \mathbf{e}_i)\widehat{\mathcal{T}}(\mathbf{I}, \mathbf{I}, \mathbf{e}_i)^\top$ for asymmetric tensor.

4.6.1 Performance of Slice-Based Initialization algorithm for symmetric tensors

Both the performance of symmetric tensor decomposition using rank-1 power method Anandkumar et al. [2014a] and that of simultaneous power method Wang and Lu [2017] will be improved using our initialization procedure. Consider a symmetric tensor with orthogonal components $\mathcal{T} = \sum_{i=1}^{R} \lambda_i \mathbf{u}_i \otimes \mathbf{u}_i \otimes \mathbf{u}_i$ where $\mathbf{u}_i \perp \mathbf{u}_j$ and $\mathbf{u}_i^\top \mathbf{u}_i = 1$. We start with a vector $\mathbf{v}^C$ which is the collection of the trace of each third mode slice of tensor $\mathcal{T}$,

i.e., the i^{th} element of vector $\mathbf{v}^C$ is $v_i^C = \sum_{l=1}^{d} \sum_{m=1}^{R} \lambda_m u_{lm} u_{lm} u_{im} \forall i \in [d]$. We then take mode-3 product of tensor $\mathcal{T}$ with the above vector $\mathbf{v}^C$. As a result, we have Lemma 4.14.

Rank-1 Power Method with deflation Anandkumar et al. [2014a] uses random unit vector initializations, and the power iteration $\mathbf{v}^{(k+1)} = \mathcal{T}(\mathbf{I}, \mathbf{v}^{(k)}, \mathbf{v}^{(k)})$ converges to the tensor eigenvector with the largest $|c_i \lambda_i|$ among $|c_1 \lambda_1|, \cdots, |c_R \lambda_R|$ where $c_i = \mathbf{v}^\top \mathbf{u}_i$. A drawback of this property is that random initialization does *not* guarantee convergence to the eigenvector with the largest eigenvalue.

Lemma 4.4 (Slice-Based Initialization improves the rank-1 power method)**.** *For each power iteration loop in rank-1 power method with deflation Anandkumar et al. [2014a] for symmetric tensors, procedure 2 guarantees recovery of the eigenvector corresponding to the largest eigenvalue.*

Slice-Based Initialization for symmetric tensors recovers the top-r subspace of the true factor $\mathbf{U}$ as descending order of λ_i^2 is the same as descending order of λ_i. Algorithm 2 uses $v_k = \text{trace}\big(\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{e}_k)\big)$ and thus $\mathbf{v} = \sum_{m=1}^{R} \lambda_m \mathbf{u}_m$. Therefore we obtain $c_i = \mathbf{v}^\top \mathbf{u}_i = \lambda_i$, and the power method converges to the eigenvector $\mathbf{u}_1$ which corresponds to the largest eigenvalue λ_1.

Likewise, **Rank-r Simultaneous Power Method** for symmetric tensors also becomes more efficient when Algorithm 2 is adopted as an initialization procedure.

Lemma 4.5 (Slice-Based Initialization improves the rank-r simultaneous power method)**.** *If algorithm 2 is used to provide an initialization for the matrix subspace iterations in Wang and Lu [2017], sampling and averaging will not be required. This can save* $O(\frac{1}{\gamma^2} \log d)$ *steps of iterations inWang and Lu [2017] where* $\gamma = \min_{1 \le i \le R} \frac{\lambda_i^2 - \lambda_{i+1}^2}{\lambda_i^2}$.

In the initialization phase of the algorithm in Wang and Lu [2017], the paper generates random Gaussian vectors $\mathbf{w}_1, \cdots, \mathbf{w}_L \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and let $\bar{\mathbf{w}} = \frac{1}{L}\sum_{l=1}^{L} \mathcal{T}(\mathbf{I}, \mathbf{w}_l, \mathbf{w}_l)$. By doing $\mathcal{T}(\mathbf{I}, \mathbf{I}, \bar{\mathbf{w}})$, Wang and Lu [2017] builds a matrix with approximately squared eigenvalues and preserved eigengaps. We improve this phase by simply obtaining vector $\mathbf{v}$ as $(\mathbf{v})_k = \text{trace}\big(\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{e}_k)\big)$ and substitute $\mathcal{T}(\mathbf{I}, \mathbf{I}, \bar{\mathbf{w}})$ by $\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{v})$.

Our Slice-Based Initialization for the symmetric case is slightly different from the asymmetric case for consideration of computational complexity (saving the multiplication of two $d \times d$ matrices). However, the asymmetric Slice-Based Initialization applies to symmetric case and allows a larger noise. Symmetric Slice-Based Initialization requires the operator norm of the noise tensor to be $O(\delta_0 \min\{\frac{\lambda_r^2 - \lambda_{r+1}^2}{4\|\lambda\|}, \frac{\lambda_r - \lambda_{r+1}}{2d^{(3/4)}}\})$, while the asymmetric Slice-Based Initialization requires the operator norm of the noise tensor to be $O(\delta_0 \min\{\frac{\lambda_r^2 - \lambda_{r+1}^2}{8\|\lambda\|}, \frac{\lambda_r - \lambda_{r+1}}{2\sqrt{d}}\})$.

4.6.2 Performance of Slice-Based Initialization algorithm for asymmetric tensor

We provide the first initialization approach for asymmetric tensors, and prove the first convergence result for asymmetric tensors. With our Slice-Based Initialization, which involves a different procedure for asymmetric tensors than for symmetric tensors, the top-r components convergence rate of asymmetric tensors matches that of symmetric tensors. Now let us consider the asymmetric tensor $\mathcal{T}$ with orthogonal components $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. We start with taking the quadratic form of each slice matrix along the third mode of the tensor, i.e., $\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{e}_i)\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{e}_i)^\top$. We obtain $\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{e}_i)\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{e}_i)^\top = \sum_{j=1}^{R} \lambda_j^2 c_{ij}^2 \mathbf{a}_j \mathbf{a}_j^\top$

which implies $\sum_{i=1}^{d} \mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{e}_i)\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{e}_i)^\top = \sum_{j=1}^{R} \lambda_j^2 \mathbf{a}_j \mathbf{a}_j^\top$ as $\mathbf{C}$ is orthonormal.

Lemma 4.6 (Preserved Component Order)**.** *Aggregated quadratic form of slices of asymmetric tensor satisfies* $\sum_{i=1}^{d} \mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{e}_i)\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{e}_i)^\top = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^\top$ *where* $\mathbf{\Lambda} = \mathsf{Diag}((\lambda_m)_{1 \le m \le R})$.

Our Slice-Based Initialization for asymmetric tensors recovers the top-r subspace of the true factors $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ as the descending order of λ_i^2 is the same as descending order of λ_i.

4.7 Robustness of the convergence result

We now extend the convergence result to noisy asymmetric tensors. For symmetric tensors, there are a number of prior efforts Sharan and Valiant [2017], Wang and Lu [2017], Anandkumar et al. [2016] showing that their decomposition algorithms are robust to noise. Such robustness depends upon restriction on tensor or structure of the noise such as low column correlations of factor matrices (in Sharan and Valiant [2017]) or symmetry of noise along with the true tensor (in Wang and Lu [2017]). We provide a robustness theorem of our algorithm under the following *bounded noise* condition.

Definition 4.3 (δ_0-bounded Noise Condition)**.** *A tensor satisfies the* δ_0*-bounded noise condition if the noise tensor is bounded in operator norm that* $\forall 1 \le r \le R$,

$$\|\mathbf{\Phi}\|_{op} \le \min\left\{ \frac{\sqrt{2}}{8} \frac{(\lambda_r - \lambda_{r+1})\epsilon}{\sqrt{r}}, \delta_0 \frac{\lambda_r^2 - \lambda_{r+1}^2}{8\|\lambda\|}, \delta_0 \frac{\lambda_r - \lambda_{r+1}}{2\sqrt{d}} \right\}.$$

Under the bounded noise model, we have the following robustness result.

Theorem 4.7 (s-ASI Convergence Guarantee)**.** *Assume the tensor $\mathcal{T}$ permits a CP decomposition form $[\![\mathbf{\Lambda}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] + \mathbf{\Phi}$ where $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are orthonormal matrices and the noise tensor $\mathbf{\Phi}$ satisfies the δ_0-bounded noise condition. For all $1 \leq r \leq R$, after $J = O(1/\log(|\frac{\lambda_r}{\lambda_{r+1}}|))$ matrix subspace iterations in procedure 3 and $O(\log(\log \frac{1}{\epsilon}))$ Alternating Subspace Iteration iterations in procedure 1, s-ASI is guaranteed to return estimated $\mathbf{\Lambda}^*$,$\mathbf{A}^*$,$\mathbf{B}^*$ and $\mathbf{C}^*$ with probability $> 1 - O(\delta_0)$. And the estimations satisfy, up to sign flip, $\|\mathbf{a}_i - \mathbf{a}_i^*\| \leq \epsilon,\ \forall 1 \leq i \leq r$. Similarly for $\mathbf{b}_i^*$ and $\mathbf{c}_i^*$ $\forall 1 \leq i \leq r$.*

The proof follows from the main convergence result 4.10 and is in Appendix 4.7.5.

Remark 1. *We make a few points below.*

1. *If the goal is to recover all components $\mathbf{A}_R, \mathbf{B}_R, \mathbf{C}_R$, then the preservation of eigenvalue order is not required. Thus the bound on the operator norm of the noise tensor can be relaxed to $O(\frac{\lambda_{\min}}{\sqrt{d}}\epsilon)$,*

2. *For the robustness theorem the worst case is considered (rather than considering the average case associated with a specific family of noise distribution), without any structural assumption. In the general case, the noise can be "malicious" if there is a sharp angle between subspace of $\mathbf{\Phi}$ and subspace of $\mathcal{T}$ for every modes.*

4.7.1 A naive initialization procedure

Based on the CP decomposition model in Equation (4.1), it is easy to see that the frontal slices shares the mode-A and mode-B singular vectors with the tensor $\mathcal{T}$, and the

k^{th} frontal slice is $\mathbf{M}_{Ck} = \mathbf{A}\mathbf{\Lambda}_{Ck}\mathbf{B}^\top$ where $\mathbf{\Lambda}_{Ck} = \begin{bmatrix} \lambda_1 c_{k1} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \lambda_R c_{kR} \end{bmatrix}$. It is natural to consider naively implementing singular value decompositions on the frontal slices to obtain estimations of $\mathbf{A}$ and $\mathbf{B}$.

Failure of Naive Initialization Consider the simpler scenario of finding a good initialization for a symmetric tensor $\mathcal{T}$ which permits the following CP decomposition

$$\mathcal{T} = \sum_{i=1}^{R} \lambda_i \mathbf{u}_i \otimes \mathbf{u}_i \otimes \mathbf{u}_i \tag{4.3}$$

Specifically we have

$$\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{v}^C) = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^\top$$

where $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_R], \Lambda = \text{diag}(\lambda_1, \cdots, \lambda_R)$. However the first method gives us a matrix without any improvement on the diagonal decomposition, i.e. $\mathbf{U}\mathbf{\Lambda}_U\mathbf{U}^\top$, where

$$\mathbf{\Lambda}_U = \text{diag}(\lambda_1 u_{k1}, \cdots, \lambda_R u_{kR})$$

For each eigenvalue of matrix $\mathbf{U}\mathbf{\Lambda}_U\mathbf{U}^\top$, it contains not only the factor of a tensor singular value which we care about, but also some unknowns from the unitary matrix. This induces trouble when one wants to recover the subspace relative to only some leading singular values of the tensor if the rank R is believed to be in a greater order of the dimension

d. Although the analogous statement in matrix subspace iteration is true almost surely (with probability one), in tensor subspace iteration we indeed need to do more work than simply taking a slice. It is highly likely that the unknown entries $u_{k1}, \cdots, u_{kR}$ permute the eigenvalues into an unfavorable sequence. Meanwhile, since $\mathbf{\Lambda}^2$ is ideally clean, we see success when we use the second method to recover the subspace relative to a few dominant singular values of a symmetric tensor.

They are all qualified in the sense that they own $\mathbf{A}$ as the left eigenspace exactly. However we can generalize this scheme to a greater extent. Frontal slicing is just a specific realization of multiplying the tensor on the third mode by a unit vector. Mode-n product of a tensor with a vector would return the collection of inner products of each mode-n fiber with the vector. The mode-3 product of tensor $\mathcal{T}$ with $\mathbf{e}_k$ will give the kth slice of $\mathcal{T}$.

4.7.2 Unreliability of symmetrization

In multi-view model, Anandkumar et al. [2012] introduced a method to symmetrize an asymmetric tensor. Here we change the notations and restate it below.

Proposition 4.8. *Let* $\mathcal{T} = \sum_{i=1}^{R} \lambda_i \mathbf{u}_i \otimes \mathbf{v}_i \otimes \mathbf{w}_i$ *have components* $\mathbf{U}, \mathbf{V}, \mathbf{W}$*, then for some vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *chosen independently, tensor*

$$\mathcal{T}(\mathcal{T}(\mathbf{b}, \mathbf{I}, \mathbf{I})^\top \mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{a})^{-1}, \mathcal{T}(\mathbf{I}, \mathbf{b}, \mathbf{I})^\top (\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{a})^\top)^{-1}, \mathbf{I}) \tag{4.4}$$

is symmetric.

Proof.

$$\mathcal{T}(\mathbf{I},\mathbf{I},\mathbf{a}) = \sum_{i=1}^{R} \lambda_i(\mathbf{w}_i^\top \mathbf{a})\mathbf{u}_i \otimes \mathbf{v}_i = \mathbf{U}\mathsf{Diag}(\lambda_i(\mathbf{w}_i^\top \mathbf{a}))\mathbf{V}^\top$$

$$(\mathcal{T}(\mathbf{I},\mathbf{I},\mathbf{a}))^{-1} = \mathbf{V}\mathsf{Diag}\big(\frac{1}{\lambda_i(\mathbf{w}_i^\top \mathbf{a})}\big)\mathbf{U}^\top$$

Similarly,

$$\mathcal{T}(\mathbf{b},\mathbf{I},\mathbf{I}) = \mathbf{V}\mathsf{Diag}(\lambda_i(\mathbf{u}_i^\top \mathbf{b}))\mathbf{W}^\top, \quad \mathcal{T}(\mathbf{I},\mathbf{b},\mathbf{I}) = \mathbf{U}\mathsf{Diag}(\lambda_i(\mathbf{v}_i^\top \mathbf{b}))\mathbf{W}^\top \tag{4.5}$$

Therefore,

$$\begin{aligned}
&\mathcal{T}(\mathcal{T}(\mathbf{b},\mathbf{I},\mathbf{I})^\top \mathcal{T}(\mathbf{I},\mathbf{I},\mathbf{a})^{-1}, \mathcal{T}(\mathbf{I},\mathbf{b},\mathbf{I})^\top (\mathcal{T}(\mathbf{I},\mathbf{I},\mathbf{a})^\top)^{-1}, \mathbf{I}) \\
&= \mathcal{T}(\mathbf{W}\mathsf{Diag}\Big(\frac{\mathbf{u}_i^\top \mathbf{b}}{\mathbf{w}_i^\top \mathbf{a}}\Big)\mathbf{U}^\top, \mathbf{W}\mathsf{Diag}\Big(\frac{\mathbf{v}_i^\top \mathbf{b}}{\mathbf{w}_i^\top \mathbf{a}}\Big)\mathbf{V}^\top, \mathbf{I}) \\
&= \sum_{i=1}^{R} \lambda_i \frac{\mathbf{u}_i^\top \mathbf{b}}{\mathbf{w}_i^\top \mathbf{a}} \frac{\mathbf{v}_i^\top \mathbf{b}}{\mathbf{w}_i^\top \mathbf{a}} \mathbf{w}_i \otimes \mathbf{w}_i \otimes \mathbf{w}_i
\end{aligned}$$

shows the symmetry. □

However, in practice the condition number for $\mathcal{T}(\mathbf{I},\mathbf{I},\mathbf{a})$ could be very large. So symmetrization using matrix inversion is not reliable since it is sensitive to noise.

Indeed, we can analyze this assuming $\mathbf{a}$ is a fixed vector. Proposition 4.9 by Jiang et al. Jiang [2006] provides a good tool for our analysis.

Proposition 4.9. *Let* $\mathbf{M}_d = (m_{ij})_{1\le i,j\le d}$*, where* m_{ij}*'s are independent standard Gaussian,* $\mathbf{X}_d = (x_{ij})_{1\le i,j\le d}$ *be the matrix obtained from performing the Gram-Schmidt pro-*

cedure on the columns of $\mathbf{M}_d$*,* $\{n_d < d : d \geq 1\}$ *be a sequence of positive integers and*

$$\epsilon_d(n) \equiv \max_{1 \leq i \leq d, 1 \leq j \leq n} \left| \sqrt{d} x_{ij} - m_{ij} \right|,$$

we then have

(1) the matrix $\mathbf{X}_d$ *is Haar invariant on the orthonormal group* $O(n)$*;*

(2) $\epsilon_d(n_d) \to 0$ *in probability, provided* $n_d = o(d/\log d)$ *as* $n \to \infty$*;*

(3) $\forall \alpha > 0$*, we have that* $\epsilon_d([d\alpha/\log d]) \to 2\sqrt{\alpha}$ *in probability as* $d \to \infty$*.*

This proposition states that for an orthonormal matrix generated by performing Gram-Schmidt procedure to standard normal matrix, , the first $o(d/\log d)$ columns, scaled by $\sqrt{d}$, asymptotically behave like a matrix with independent standard Gaussian entries and this is the largest order for the number of columns we can approximate simultaneously.

The condition number of matrix $\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{a})$ is

$$\mathcal{K}(\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{a})) = \frac{\max\limits_{1 \leq i \leq R} |\lambda_i \mathbf{w}_i^\top \mathbf{a}|}{\min\limits_{1 \leq i \leq R} |\lambda_i \mathbf{w}_i^\top \mathbf{a}|}, \tag{4.6}$$

which is nondecreasing as the rank of tensor R increases. So we can indeed assume $R = o(d/\log d)$ and study the badness of condition number for such $\mathbf{W}$'s as worse cases.

Remark 2. *We treat* $\mathbf{W}$ *as the left* $d \times R$ *sub-block of some orthonormal matrix. Thus by assuming* $R = o(d/\log d)$*,* $\sqrt{d}\mathbf{W}$ *could be approximated by a matrix of i.i.d.* $\mathcal{N}(0, 1)$

variables when d is large, which is common in practice.

Since condition number $\mathcal{K}$ is taking ratio, without loss of gernerality we can let $\|\mathbf{a}\| = 1$. Then,

$$\mathcal{K}(\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{a})) = \frac{\max\limits_{1\le i\le R} |\lambda_i(\sqrt{d}\mathbf{w}_i)^\top \mathbf{a}|}{\min\limits_{1\le i\le R} |\lambda_i(\sqrt{d}\mathbf{w}_i)^\top \mathbf{a}|}. \tag{4.7}$$

For $1 \le i \le R$, $\lambda_i(\sqrt{d}\mathbf{w}_i)^\top \mathbf{a}$ are independent to each other and approximately has distribution $\mathcal{N}(0, \lambda_i^2)$. So the condition number is approximately the ratio between maximum and minimum of absolute value of $\mathcal{N}(\mathbf{0}, \mathrm{Diag}(\lambda_i^2))$. One can imgine if the tensor has one or more small singular values then it is highly likely for the condition number to be high.

4.7.3 Procedure 1 noiseless convergence result

4.7.3.1 Conditional simultaneous convergence

Theorem 4.10 (Main Convergence)**.** *Using the initialization procedure 2, Denote the recovered tensor as $\mathcal{T}^* = [\![\mathbf{\Lambda}^*; \mathbf{A}^*, \mathbf{B}^*, \mathbf{C}^*]\!]$ after $J = O(\log(C)/\log(|\frac{\lambda_r}{\lambda_{r+1}}|))$ iterations in initialization procedure 2 and $K = O(\log(\log\frac{1}{\epsilon}))$ iterations in main procedure1 applied on $\mathcal{T}$, $\forall \epsilon > 0$. We have*

$$\|\mathcal{T}^* - \mathcal{T}\|_s \le \epsilon.$$

To prove the main convergence result, just combine all of the rest results together.

Lemma 4.11. *Let $\mathbf{Q}_{\mathbf{A}_r}^{(0)}, \mathbf{Q}_{\mathbf{B}_r}^{(0)}, \mathbf{Q}_{\mathbf{C}_r}^{(0)}, \forall r \in \{1, 2, \cdots, R\}$, be $d \times r$ orthonormal initializa-*

tion matrices for the specified subspace iteration. Then after K iterations, we have

$$t_{A_{(r)}}^{(K)} \leq \left(\frac{\lambda_{r+1}}{\lambda_{(r)}}\right)^{2^K-1} \left(t_{A_{(r)}}^{(0)} t_{B_{(r)}}^{(0)} t_{C_{(r)}}^{(0)}\right)^{\frac{2^K}{3}} \left[\frac{\left(t_{A_{(r)}}^{(0)}\right)^2}{t_{B_{(r)}}^{(0)} t_{C_{(r)}}^{(0)}}\right]^{\frac{(-1)^K}{3}}, \quad \forall K \geq 1.$$

where $t_{A_{(r)}}^{(k)} = \tan\left(\mathbf{A}_r, \mathbf{Q}_{\mathbf{A}_r}^{(k)}\right)$, $t_{B_{(r)}}^{(k)} = \tan\left(\mathbf{B}_{(r)}, \mathbf{Q}_{\mathbf{B}_r}^{(k)}\right)$, $t_{C_{(r)}}^{(k)} = \tan\left(\mathbf{C}_{(r)}, \mathbf{Q}_{\mathbf{C}_r}^{(k)}\right)$, $\forall k \geq 0$. *Similarly for* $\mathbf{B}_{(r)}$ *and* $\mathbf{C}_{(r)}$.

The proof is in Appendix 4.7.3.2.

Remark 3. *Given that the initialization matrices* $\mathbf{Q}_{\mathbf{A}_r}^{(0)}, \mathbf{Q}_{\mathbf{B}_r}^{(0)}, \mathbf{Q}_{\mathbf{C}_r}^{(0)}$ *satisfy the* r*-sufficient initialization condition, the angles between approximate subspaces and true spaces would decrease with a quadratic rate. Therefore, only* $K = O(\log(\log \frac{1}{\epsilon}))$ *number of iterations is needed to achieve* $\tan(\mathbf{A}_r, \mathbf{Q}_{\mathbf{A}_r}^{(K)}) \leq \epsilon$.

The following result shows that if we have the angle of subspaces small enough, column vectors of the approximate matrix converges simultaneously to the true vectors of true tensor component at the same position.

Lemma 4.12 (Simultaneous Convergence)**.** *For any* $r \in \{1, 2, \cdots, R\}$, *if*

$$\tan(\mathbf{A}_r, \mathbf{Q}_{\mathbf{A}_r}) \leq \epsilon \tag{4.8}$$

for some $d \times r$ *matrix* $\mathbf{Q}_{\mathbf{A}_r} = [\mathbf{q}_1, \cdots, \mathbf{q}_r]$, *then*

$$\|\mathbf{q}_i - \mathbf{a}_i\|^2 \leq 2\epsilon, \quad \forall 1 \leq i \leq r.$$

Similarly for $\mathbf{B}_{(r)}$ *and* $\mathbf{C}_{(r)}$.

The proof is in Appendix 4.7.3.3.

4.7.3.2 Proof for lemma 4.11

Proof. We only prove the result for the order of A. The proofs for the other two orders are the same.

For rank-R tensor $\mathcal{T} = [\![\mathbf{\Lambda}; \mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \equiv \sum_{i=1}^{R} \lambda_i \mathbf{a}_i \otimes \mathbf{b}_i \otimes \mathbf{c}_i$, its mode-1 matricization $\mathcal{T}_{(1)} = \mathbf{A}\mathbf{\Lambda}(\mathbf{C} \odot \mathbf{B})^\top$. So in each iteration,

$$\begin{aligned}\mathbf{Q}_{\mathbf{A}_r}^{(k+1)}\mathbf{R}_{\mathbf{A}_r}^{(k+1)} &= \mathcal{T}_{(1)}(\mathbf{Q}_{\mathbf{C}_r}^{(k)} \odot \mathbf{Q}_{\mathbf{B}_r}^{(k)}) = \mathbf{A}\mathbf{\Lambda}(\mathbf{C} \odot \mathbf{B})^\top(\mathbf{Q}_{\mathbf{C}_r}^{(k)} \odot \mathbf{Q}_{\mathbf{B}_r}^{(k)}) \\ &= \mathbf{A}\mathbf{\Lambda}(\mathbf{C}^\top\mathbf{Q}_{\mathbf{C}_r}^{(k)}) * (\mathbf{B}^\top\mathbf{Q}_{\mathbf{B}_r}^{(k)})\end{aligned}$$

by property of Hadamard product and Khatri-Rao product Liu and Trenkler [2008], Kolda and Bader [2009].

We can expand matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ to be a basis for $\mathbb{R}^d$, and we can for example for $\mathbf{A}_r$, let $\mathbf{A}_r^c$ be the matrix consisted of the rest $(d - r)$ columns in the expanded matrix. Now the column space of $\mathbf{A}_r^c$ is just the complement space of column space of $\mathbf{A}_r$ in $\mathbb{R}^d$. And $\begin{bmatrix}\mathbf{A}_r & \mathbf{A}_r^c\end{bmatrix}$ is a $d \times d$ orthonormal matrix.

With that notation, we have for $0 \leq k \leq K$,

$$\begin{aligned}\mathbf{A}_r^\top\mathbf{Q}_{\mathbf{A}_r}^{(k+1)}\mathbf{R}_{\mathbf{A}_r}^{(k+1)} &= \begin{bmatrix}\mathbf{I}_r & \mathbf{0}_{r\times(R-r)}\end{bmatrix}\mathbf{\Lambda}\big(\mathbf{C}^\top\mathbf{Q}_{\mathbf{C}_r}^{(k)}\big) * \big(\mathbf{B}^\top\mathbf{Q}_{\mathbf{B}_r}^{(k)}\big) \\ \mathbf{A}_r^{c\top}\mathbf{Q}_{\mathbf{A}_r}^{(k+1)}\mathbf{R}_{\mathbf{A}_r}^{(k+1)} &= \begin{bmatrix}\mathbf{0}_{(R-r)\times r} & \mathbf{I}_{(R-r)\times(R-r)} \\ \mathbf{0}_{(d-R)\times r} & \mathbf{0}_{(d-R)\times(R-r)}\end{bmatrix}\mathbf{\Lambda}\big(\mathbf{C}^\top\mathbf{Q}_{\mathbf{C}_r}^{(k)}\big) * \big(\mathbf{B}^\top\mathbf{Q}_{\mathbf{B}_r}^{(k)}\big).\end{aligned}$$

Now fix k and focus on a single iteratoin step,

$$
\begin{aligned}
t_{A_r}^{(k+1)} &= \tan(\mathbf{A}_r, \mathbf{Q}_{\mathbf{A}_r}^{(k+1)}) = \frac{\sin(\mathbf{A}_r, \mathbf{Q}_{\mathbf{A}_r}^{(k+1)})}{\cos(\mathbf{A}_r, \mathbf{Q}_{\mathbf{A}_r}^{(k+1)})} = \frac{\sigma_{\max}(\mathbf{A}_r^{c\top}\mathbf{Q}_{\mathbf{A}_r}^{(k+1)})}{\sigma_{\min}(\mathbf{A}_r^{\top}\mathbf{Q}_{\mathbf{A}_r}^{(k+1)})} \\
&= \left\|\mathbf{A}_r^{c\top}\mathbf{Q}_{\mathbf{A}_r}^{(k+1)}\right\|_s \left\|\left(\mathbf{A}_r^{\top}\mathbf{Q}_{\mathbf{A}_r}^{(k+1)}\right)^{-1}\right\|_s \\
&= \left\|\mathbf{A}_r^{c\top}\mathbf{Q}_{\mathbf{A}_r}^{(k+1)}\left(\mathbf{A}_r^{\top}\mathbf{Q}_{\mathbf{A}_r}^{(k+1)}\right)^{-1}\right\|_s \\
&= \left\|\mathbf{A}_r^{c\top}\mathbf{Q}_{\mathbf{A}_r}^{(k+1)}\mathbf{R}_{\mathbf{A}_r}^{(k+1)}\left(\mathbf{A}_r^{\top}\mathbf{Q}_{\mathbf{A}_r}^{(k+1)}\mathbf{R}_{\mathbf{A}_r}^{(k+1)}\right)^{-1}\right\|_s \\
&\leq \frac{\sigma_{\max}\left(\mathbf{A}_r^{c\top}\mathbf{Q}_{\mathbf{A}_r}^{(k+1)}\mathbf{R}_{\mathbf{A}_r}^{(k+1)}\right)}{\sigma_{\min}\left(\mathbf{A}_r^{\top}\mathbf{Q}_{\mathbf{A}_r}^{(k+1)}\mathbf{R}_{\mathbf{A}_r}^{(k+1)}\right)} \\
&\leq \frac{\lambda_{r+1}\sigma_{\max}\left[\left(\mathbf{C}_{(r)}^{c\top}\mathbf{Q}_{\mathbf{C}_r}^{(k)}\right) * \left(\mathbf{B}_r^{c\top}\mathbf{Q}_{\mathbf{B}_r}^{(k)}\right)\right]}{\lambda_r\sigma_{\min}\left[\left(\mathbf{C}_r^{c\top}\mathbf{Q}_{\mathbf{C}_r}^{(k)}\right) * \left(\mathbf{B}_r^{c\top}\mathbf{Q}_{\mathbf{B}_r}^{(k)}\right)\right]}
\end{aligned}
$$

For Hadamard product, $\sigma_{\max}(\mathbf{M}_1 * \mathbf{M}_2) \leq \sigma_{\max}(\mathbf{M}_1)\sigma_{\max}(\mathbf{M}_2)$

and $\sigma_{\min}(\mathbf{M}_1 * \mathbf{M}_2) \geq \sigma_{\min}(\mathbf{M}_1)\sigma_{\min}(\mathbf{M}_2)$see Liu and Trenkler [2008]

$$
\begin{aligned}
&\leq \frac{\lambda_{r+1}}{\lambda_r}\frac{\sigma_{\max}\left(\mathbf{C}_r^{c\top}\mathbf{Q}_{\mathbf{C}_r}^{(k)}\right)}{\sigma_{\min}\left(\mathbf{C}_r^{c\top}\mathbf{Q}_{\mathbf{C}_r}^{(k)}\right)}\frac{\sigma_{\max}\left(\mathbf{B}_r^{c\top}\mathbf{Q}_{\mathbf{B}_r}^{(k)}\right)}{\sigma_{\min}\left(\mathbf{B}_r^{c\top}\mathbf{Q}_{\mathbf{B}_r}^{(k)}\right)} \\
&= \frac{\lambda_{r+1}}{\lambda_r} \cdot \tan\left(\mathbf{B}_r, \mathbf{Q}_{\mathbf{B}_r}^{(k)}\right) \cdot \tan\left(\mathbf{C}_r, \mathbf{Q}_{\mathbf{C}_r}^{(k)}\right)
\end{aligned}
$$

Therefore we get $\forall 0 \leq k \leq K$,

$$
t_{A_r}^{(k+1)} \leq \frac{\lambda_{r+1}}{\lambda_r} t_{B_r}^{(k)} t_{C_r}^{(k)}.
$$

And similarly,

$$
t_{B_r}^{(k+1)} \leq \frac{\lambda_{r+1}}{\lambda_r} t_{A_r}^{(k)} t_{C_r}^{(k)},
$$

$$t_{C_r}^{(k+1)} \leq \frac{\lambda_{r+1}}{\lambda_r} t_{A_r}^{(k)} t_{B_r}^{(k)}.$$

Sequentially,

$$\begin{aligned} t_{A_r}^{(K+1)} &\leq \frac{\lambda_{r+1}}{\lambda_r} t_{B_r}^{(K)} t_{C_r}^{(K)} \leq \Big(\frac{\lambda_{r+1}}{\lambda_r}\Big)^3 (t_{A_r}^{(K-1)})^2 t_{B_r}^{(K-1)} t_{C_r}^{(K-1)} \\ &\leq \cdots \leq \Big(\frac{\lambda_{r+1}}{\lambda_r}\Big)^{1+2m} \Big(\prod_{i=1}^{m} (t_{A_r}^{(K-i)})^2\Big) t_{B_r}^{(K-m)} t_{C_r}^{(K-m)} \quad \forall m = 1, 2, \ldots, K \end{aligned}$$

Easy to see that all historical tangents of principal angle in approximation for $\mathbf{A}_r$ appear in the upper bound for the tangent-measured approximation distance after a new iteration. So in order to solve for the explicit upper bounds, we can assume the form of the upper bounds has a recursive formula for each exponents. Specifically, assume for some sequences u_K, a_K, b_K, we can conclude

$$t_{A_r}^{K+1} \leq \Big(\frac{\lambda_{r+1}}{\lambda_r}\Big)^{u_{K+1}} \big(t_{A_r}^{(0)}\big)^{a_{K+1}} \big(t_{B_r}^{(0)} t_{A_r}^{(0)}\big)^{b_{K+1}}$$

On the other hand, for fixed $K \geq 1$,

$$\begin{aligned} t_{A_r}^{K+1} &\leq \Big(\frac{\lambda_{r+1}}{\lambda_r}\Big)^{1+2K} \Big(\prod_{i=1}^{K} (t_{A_r}^{(K-i)})^2\Big) t_{B_r}^{(0)} t_{C_r}^{(0)} \\ &\leq \Big(\frac{\lambda_{r+1}}{\lambda_r}\Big)^{1+2K} \prod_{i=1}^{K} \Big[\Big(\frac{\lambda_{r+1}}{\lambda_r}\Big)^{u_{K-i}} \big(t_{A_r}^{(0)}\big)^{a_{K-i}} \big(t_{B_r}^{(0)} t_{A_r}^{(0)}\big)^{b_{K-i}}\Big]^2 \cdot t_{B_r}^{(0)} t_{C_r}^{(0)} \\ &= \Big(\frac{\lambda_{r+1}}{\lambda_r}\Big)^{1+2K+2\sum_{i=1}^{K} u_{K-i}} \big(t_{A_r}^{(0)}\big)^{2\sum_{i=1}^{K} a_{K-i}} \big(t_{B_r}^{(0)} t_{A_r}^{(0)}\big)^{1+2\sum_{i=1}^{K} b_{K-i}} \end{aligned}$$

Now we have gained the recursive formulas for sequence on exponents in the upper

bound

$$u_{K+1} = 1 + 2K + 2\sum_{i=1}^{K} u_{K-i}$$
$$a_{K+1} = 2\sum_{i=1}^{K} a_{K-i}$$
$$b_{K+1} = 1 + 2\sum_{i=1}^{K} b_{K-i}.$$

The formula system works on when $K \geq 1$, so we can check the upper bounds for several initial iterations.

For $K = 0$,

$$t_{A_r}^{(1)} \leq \frac{\lambda_{r+1}}{\lambda_r} t_{B_r}^{(0)} t_{C_r}^{(0)}$$

For $K = 1$,

$$t_{A_r}^{(2)} \leq \left(\frac{\lambda_{r+1}}{\lambda_r}\right)^3 \left(t_{A_r}^{(0)}\right)^2 t_{B_r}^{(0)} t_{C_r}^{(0)}$$

For $K = 2$,

$$t_{A_r}^{(3)} \leq \left(\frac{\lambda_{r+1}}{\lambda_r}\right)^7 \left(t_{A_r}^{(0)}\right)^2 \left(t_{B_r}^{(0)} t_{C_r}^{(0)}\right)^3$$

We have

$$u_0 = 0, u_1 = 1, u_2 = 3, u_3 = 7, u_4 = 15, \ldots$$

$$a_0 = 1, a_1 = 0, a_2 = 2, a_3 = 2, a_4 = 6, \ldots$$

$$b_0 = 0, b_1 = 1, b_2 = 1, b_3 = 3, b_4 = 5, \ldots$$

One can solve and check the general formula for these sequences

$$u_K = 2^K - 1, \quad a_K = \frac{2}{3}(2^{K-1} + (-1)^K), \quad b_K = \frac{1}{3}(2^K + (-1)^{K-1}), \quad \forall K \geq 1.$$

In conclusion,

$$t_{A_r}^{(K)} \leq \left(\frac{\lambda_{r+1}}{\lambda_r}\right)^{2^K-1} \left(t_{A_r}^{(0)} t_{B_r}^{(0)} t_{C_r}^{(0)}\right)^{\frac{2^K}{3}} \left[\frac{\left(t_{A_r}^{(0)}\right)^2}{t_{B_r}^{(0)} t_{C_r}^{(0)}}\right]^{\frac{(-1)^K}{3}}, \quad \forall K \geq 1.$$

The proofs of upper bounds for $\mathbf{B}_r$ and $\mathbf{C}_r$ are the same.

□

4.7.3.3 Proof for lemma 4.12

Proof. First, we denote $\mathbf{Q}_i := [\mathbf{q}_1, \cdots, \mathbf{q}_i]$ only in this proof. Then

$$\begin{aligned}
\tan(\mathbf{A}_{r-1}, \mathbf{Q}_{r-1}) &= \frac{\sqrt{1 - \sigma_{\min}^2(\mathbf{A}_{r-1}^\top \mathbf{Q}_{r-1})}}{\sigma_{\min}(\mathbf{A}_{r-1}^\top \mathbf{Q}_{r-1})} \\
&= \sqrt{\frac{1}{\sigma_{\min}^2(\mathbf{A}_{r-1}^\top \mathbf{Q}_{r-1})} - 1} \\
&\qquad \text{by Cauchy interlacing theorem} \\
&\leq \sqrt{\frac{1}{\sigma_{\min}^2(\mathbf{A}_{(r)}^\top \mathbf{Q}_{(r)})} - 1} \\
&= \tan(\mathbf{A}_{(r)}, \mathbf{Q}_{(r)})
\end{aligned}$$

Inductively, $\forall 1 \leq i \leq r,\ \tan(\mathbf{A}_i, \mathbf{Q}_i) \leq \epsilon$. Then $\forall 2 \leq i \leq r$,

$$\begin{aligned}
\cos^2(\mathbf{A}_i, \mathbf{Q}_i) &= \min_{\mathbf{y} \in \mathbb{R}^i} \frac{\|\mathbf{Q}_i^\top \mathbf{A}_i \mathbf{y}\|^2}{\|\mathbf{A}_i \mathbf{y}\|^2} \\
&\leq \|\mathbf{Q}_i^\top \mathbf{a}_i\|^2 \quad \text{as letting } \mathbf{y} \text{ to be } [0, \cdots, 0, 1]^\top \\
&= \|\mathbf{Q}_{i-1}^\top \mathbf{a}_i\|^2 + (\mathbf{q}_i^\top \mathbf{a}_i)^2 \\
&\leq \sin^2(\mathbf{A}_{i-1}, \mathbf{Q}_{i-1}) + (\mathbf{q}_i^\top \mathbf{a}_i)^2,
\end{aligned}$$

since $\mathbf{a}_i \in \mathscr{C}(\mathbf{A}_{i-1})^\perp$, the complement space of column space of $\mathbf{A}_{i-1}$, and

$$\begin{aligned}
(\mathbf{q}_i^\top \mathbf{a}_i)^2 &\geq \frac{1}{1 + \tan^2(\mathbf{A}_i, \mathbf{Q}_i)} - \frac{\tan^2(\mathbf{A}_{i-1}, \mathbf{Q}_{i-1})}{1 + \tan^2(\mathbf{A}_{i-1}, \mathbf{Q}_{i-1})} \\
&\geq \frac{1}{1+\epsilon^2} - 1 + \frac{1}{1+\epsilon^2} = 1 - \frac{2\epsilon^2}{1+\epsilon^2} \geq 1 - 2\epsilon^2.
\end{aligned}$$

For $i = 1$,

$$\cos^2(\mathbf{A}_1, \mathbf{Q}_1) = (\mathbf{q}_1^\top \mathbf{a}_1)^2 = \frac{1}{1 + \tan^2(\mathbf{A}_1, \mathbf{Q}_1)} \geq \frac{1}{1+\epsilon^2} \geq 1 - 2\epsilon^2.$$

To conclude, $\|\mathbf{q}_i - \mathbf{a}_i\|^2 = 2 - 2\mathbf{q}_i^\top \mathbf{a}_i \leq 2\epsilon, \quad \forall 1 \leq i \leq r$. And the proofs for $\mathbf{B}_r$ and $\mathbf{C}_r$ are the same.

□

4.7.3.4 Lemma 4.13 and proof

Lemma 4.13. *Let $\mathbf{U}_p, \mathbf{V}_p \in \mathbb{R}^{d \times p}$ respectively be the orthonormal complex matrix whose column space is the left and right invariant subspace corresponding to the dominant p*

eigenvalues of $\mathbf{M} \in \mathbb{R}^{d\times d}$. *Assume for fixed initialization* $\mathbf{Q}^{(0)}$, $\mathbf{V}_p^\top \mathbf{Q}^{(0)}$ *has full rank. Then after* $\forall k \geq 1$ *steps (independent of* ϵ*) of matrix subspace iteration* $\mathbf{Q}^{(k)}\mathbf{R}^{(k)} \leftarrow QR\left(\mathbf{M}\mathbf{Q}^{(k-1)}\right)$, *we obtain* $\tan(\mathbf{U}_p, \mathbf{Q}^{(k)}) \leq C \cdot \left|\frac{\sigma_{p+1}(\mathbf{M})}{\sigma_p(\mathbf{M})}\right|^k$ *for a finite constant* C, *where* $\sigma_p(\cdot)$ *denotes the* p^{th} *singular value.*

Proof. Since $\mathbf{A}$ is orthogonal in the way $\mathbf{A}\mathbf{A}^* = \mathbf{A}^*\mathbf{A}$, $\mathbf{A}$ is a normal matrix. So its Schur decomposition and eigendecomposition coincides to $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^*$. Here $\mathbf{P}\mathbf{P}^* = \mathbf{P}^*\mathbf{P} = \mathbf{I}$. $\mathbf{D}$ is a diagonal matrix with all eigenvalues of $\mathbf{A}$ on diagonal and without loss of generality we can permutate them to be in a decreasing order, i.e. $\mathbf{D} = diag(\lambda_1, \cdots, \lambda_p, \lambda_{p+1}, \cdots, \lambda_d)$. We can furthermore denote $\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{bmatrix}$, where $\mathbf{D}_1$ contains eigenvalues up to λ_p and $\mathbf{D}_2$ contains eigenvalues λ_{p+1} to λ_d.

Inspired by Arbenz et al. [2012], without making any restriction to the matrix to initialize the algorithm, we can assume the iterations take place in the space of $\{\mathbf{P}\mathbf{Q}\}$ without loss of generality because $\mathbf{P}$ is invertible. Then we notice that for the iteration formula, it becomes

$$\mathbf{P}\mathbf{Q}^{(\mathbf{k})}\mathbf{R}^{(\mathbf{k})} := \mathbf{A}\mathbf{P}\mathbf{Q}^{(\mathbf{k-1})}$$

$$\mathbf{Q}^{(\mathbf{k})}\mathbf{R}^{(\mathbf{k})} := \mathbf{P}^*\mathbf{A}\mathbf{P}\mathbf{Q}^{(\mathbf{k-1})}$$

$$\mathbf{Q}^{(\mathbf{k})}\mathbf{R}^{(\mathbf{k})} := \mathbf{D}\mathbf{Q}^{(\mathbf{k-1})}$$

So analytically, the convergence for an arbitrary matrix is the same to the convergence for the diagonal matrix formed from the eigenvalues of that matrix. And the left invariant eigenvector subspace for $\mathbf{D}$ is nothing but $\mathbf{E}_p = [\mathbf{e_1}, \cdots, \mathbf{e_p}]$. Imgine now $\mathbf{Q}^{(0)}$ is prepared to run the algorithm for $\mathbf{D}$, next we will show the subspace of $\mathbf{Q}^{(k)}$'s will

converge to column space of $\mathbf{E}_p$.

First, partition $\mathbf{Q}^{(k)}$ to $\begin{bmatrix} \mathbf{Q}_1^{(k)} \\ \mathbf{Q}_2^{(k)} \end{bmatrix}$ such that $\mathbf{Q}_1^{(k)} \in \mathbb{C}^{p\times p}$. $\mathbf{D}_1 \in \mathbb{C}^{p\times p}$ is invertible because of the eigenvalue gap. By the assumption that $\mathbf{V}_p^*\mathbf{Q}$ has full rank, here we have $\mathbf{Q}_1^{(0)}$ has full rank and thus invertible. $\mathbf{Q}_1^{(k)}$ is therefore invertible.

Notice that inductively,

$$\mathbf{Q^{(k)}R^{(k)}} = \mathbf{DQ^{(k-1)}}$$

$$\mathbf{Q^{(k)}R^{(k)}R^{(k-1)}} = \mathbf{DQ^{(k-1)}R^{(k-1)}} = \mathbf{D^2Q^{(k-2)}}$$

$$\mathbf{Q^{(k)}R^{(k)}R^{(k-1)}\cdots R^{(1)}} = \mathbf{D^kQ^{(0)}} = \mathbf{Q^{(k)}R}$$

for some upper-triangular matrix $\mathbf{R}$. Then

$$\mathbf{Q}^{(k)}\mathbf{R} = \mathbf{D}^k\mathbf{Q}^{(0)} = \begin{bmatrix} \mathbf{D}_1^k\mathbf{Q}_1^{(0)} \\ \mathbf{D}_2^k\mathbf{Q}_2^{(0)} \end{bmatrix}.$$

$$\mathbf{Q}^{(k)} = \begin{bmatrix} \mathbf{D}_1^k\mathbf{Q}_1^{(0)}\mathbf{R}^{-1} \\ \mathbf{D}_2^k\mathbf{Q}_2^{(0)}\mathbf{R}^{-1} \end{bmatrix}$$

To study tangent, first look at

$$\begin{aligned}\sin(\mathbf{E}_p, \mathbf{Q}^{(k)}) &= \| \begin{bmatrix} \mathbf{0} & \mathbf{I}_{d-p} \end{bmatrix}^\top \mathbf{Q}^{(k)} \|_s = \|\mathbf{D}_2^k \mathbf{Q}_2^{(0)} \mathbf{R}^{-1}\|_s \\ &= \frac{\|\mathbf{D}_2^k \mathbf{Q}_2^{(0)} \mathbf{R}^{-1} (\mathbf{D}_1^k \mathbf{Q}_1^{(0)} \mathbf{R}^{-1})^{-1}\|_s}{\sqrt{1 + \|\mathbf{D}_2^k \mathbf{Q}_2^{(0)} \mathbf{R}^{-1} (\mathbf{D}_1^k \mathbf{Q}_1^{(0)} \mathbf{R}^{-1})^{-1}\|_s^2}} \\ &\quad \text{Denote } \mathbf{M}^{(k)} := \mathbf{D}_2^k \mathbf{Q}_2^{(0)} \big(\mathbf{Q}_1^{(0)}\big)^{-1} \mathbf{D}_1^{-k} \\ &= \frac{\|\mathbf{M}^{(k)}\|_s}{\sqrt{1 + \|\mathbf{M}^{(k)}\|_s^2}}.\end{aligned}$$

Correspondingly,

$$\cos(\mathbf{E}_p, \mathbf{Q}^{(k)}) = \frac{1}{\sqrt{1 + \|\mathbf{M}^{(k)}\|_s^2}}$$

Since spectral radius $\rho(\mathbf{D}_1^{-1}) = |\lambda_p|^{-1}, \rho(\mathbf{D}_2) = |\lambda_{p+1}|$, for any $\epsilon > 0$, there exists a norm $\|\cdot\|_{(1)}$ such that $\|\mathbf{D}_1^{-1}\|_{(1)} \leq |\lambda_p|^{-1} + \epsilon$, and another norm $\|\cdot\|_{(2)}$ such that $\|\mathbf{D}_2\|_{(2)} \leq |\lambda_{p+1}| + \epsilon$. By equivalence of norms, There exists constants $C_1, C_2 < \infty$ such that $\|\mathbf{M}\|_s \leq C_1 \|\mathbf{M}\|_{(1)}$ and $\|\mathbf{M}\|_s \leq C_2 \|\mathbf{M}\|_{(2)}$ for any matrix $\mathbf{M}$.

As a consequence,

$$\begin{aligned}\tan(\mathbf{E}_p, \mathbf{Q}^{(k)}) &= \|\mathbf{M}^{(k)}\|_s \leq \|\mathbf{D}_1^k\|_s \|\mathbf{M}^{(0)}\|_s \|\mathbf{D}_2^{-k}\|_s \\ &\leq C_1 C_2 \|\mathbf{D}_1^k\|_{(1)} \|\mathbf{M}^{(0)}\|_s \|\mathbf{D}_2^{-k}\|_{(2)} \\ &\leq C_1 C_2 \tan(\mathbf{E}_p, \mathbf{Q}^{(0)}) \|\mathbf{D}_1\|_{(1)}^k \|\mathbf{D}_2^{-1}\|_{(2)}^k \\ &\leq C \Big(\big(|\lambda_{p+1}| + \epsilon\big) \big(\frac{1}{|\lambda_p|} + \epsilon\big) \Big)^k\end{aligned}$$

for some constant C after an initialization is chosen and fixed.

Let ϵ_0 be $(|\lambda_{p+1}| + \frac{1}{|\lambda_p|} + \epsilon)\epsilon$, then equivalently,

$$\tan(\mathbf{E}_p, \mathbf{Q}^{(k)}) \leq C\Big(\Big|\frac{\lambda_{p+1}}{\lambda_p}\Big| + \epsilon_0\Big)^k, \quad \forall \epsilon_0 > 0.$$

This shows the convergence of subspace iteration algorithm on recovering the left eigenspace of a matrix in complex diagonal orthonormal matrix space with a specific eigenvalue gap. By the analytical equivalence dicussed before, we have identical convergence on recovering the left eigenspace of an arbitrary orthonormal matrix. In this way, equivalently, if $\mathbf{Q}^{(0)}$ is for this algorithm on $\mathbf{A}$,

$$\tan(\mathbf{U}_p, \mathbf{Q}^{(k)}) \leq C\Big(\Big|\frac{\lambda_{p+1}}{\lambda_p}\Big| + \epsilon_0\Big)^k, \quad \forall \epsilon_0 > 0.$$

By taking infimum on ϵ_0, it becomes

$$\tan(\mathbf{U}_p, \mathbf{Q}^{(k)}) \leq C \cdot \Big|\frac{\lambda_{p+1}}{\lambda_p}\Big|^k$$

□

Remark 4. *The condition that $\mathbf{V}_p^\top \mathbf{Q}$ has full rank assumed in lemma 4.13 is satisfied almost surely (with probability 1).*

Proof. As a common procedure, to generate a random $(d \times r)$-sized orthonormal matrix, one could first generate a matrix of r columns sampled i.i.d. from d-dimensional standard normal distribution, and then perform Gram-Schmidt algorithm on columns. Consider

Gram-Schmidt algorithm as a mapping. Then under such mapping, the pre-image of a orthonormal matrix $[\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_r]$ is $[s_1\mathbf{q}_1, s_{21}\mathbf{q}_1 + s_{22}\mathbf{q}_2, \cdots, s_{r1}\mathbf{q}_1 + \cdots + s_{rr}\mathbf{q}_r]$, for some constants $s_1, s_{21}, \cdots, s_{rr} \in \mathbb{R}$. The columns of the pre-image (sampled from i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$) belong to a subspace in $\mathbb{R}^d$.

The condition that $\mathbf{V}_p^\top \mathbf{Q}$ has full rank is equivalent to the condition that there exists at least one column of $\mathbf{Q}$ that is in the complement of column space of $\mathbf{V}_p$ in $\mathbb{R}^d$. So as long as the column space of $\mathbf{V}_p$ is not the whole $\mathbb{R}^d$, in order to make $\mathbf{V}_p^\top \mathbf{Q}$ not a full-rank matrix, at least one column of the random normal matrix has to take place in a proper subspace in $\mathbb{R}^d$. The multi-variate normal distribution is also a finite measure on $\mathbb{R}^d$. Therefore the measure of that proper subspace (i.e. the probability that we fail to have a full-rank $\mathbf{V}_p^\top \mathbf{Q}$) is zero. □

4.7.4 Lemma 4.14 and proof

Lemma 4.14. *Mode-3 product of symmetric tensor $\mathcal{T}$ with vector $\mathbf{v}^C$ has the form $\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{v}^C) = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^\top$ where $\mathbf{\Lambda} = \mathsf{Diag}((\lambda_m)_{1 \le m \le R}), \mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_R]$.*

Proof. We will prove a more general case for asymmetric tensor. $\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{v}^C)$ is a matrix.

The (i, j)th entry of the matrix would be

$$
\begin{aligned}
[\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{v}^C)]_{ij} &= \sum_{k=1}^{d} \Big(\sum_{l=1}^{d} \sum_{m_1=1}^{R} \lambda_{m_1} a_{lm_1} b_{lm_1} c_{km_1} \Big) \cdot \Big(\sum_{m_2=1}^{R} \lambda_{m_2} a_{im_2} b_{jm_2} c_{km_2} \Big) \\
&= \sum_{m_1, m_2=1}^{R} \sum_{l=1}^{d} \lambda_{m_1} \lambda_{m_2} a_{lm_1} a_{im_2} b_{lm_1} b_{jm_2} \sum_{k=1}^{d} c_{km_1} c_{km_2} \\
&\quad \text{Because} \sum_{k=1}^{d} c_{km_1} c_{km_2} = \begin{cases} = 0 & \text{if } m_1 \neq m_2 \\ \\ = 1 & \text{if } m_1 = m_2 \end{cases}. \\
&= \sum_{m=1}^{R} \Big(\lambda_m^2 \sum_{l=1}^{d} a_{lm} b_{lm} \Big) a_{im} b_{jm} \\
&= \sum_{m=1}^{R} (\lambda_m^2 \mathbf{a}_m^\top \mathbf{b}_m) a_{im} b_{jm}.
\end{aligned}
$$

The symmetric tensor proof is trivial after achieving the above argument. □

4.7.5 Robustness of our algorithm under noise

Let $\mathcal{T}$ be the true tensor, $\widehat{\mathcal{T}} = \mathcal{T} + \Phi$ be the observed noisy tensor, where Φ is the noise. Let $\mathbf{M}$ and $\widehat{\mathbf{M}}$ be the matrix prepared from $\mathcal{T}$ and $\widehat{\mathcal{T}}$ by Procedure 2 for matrix subspace iteration.

4.7.6 Perturbation bounds

Lemma 4.15 (Perturbation in slice-based initialization step)**.**

$$
\|\widehat{\mathbf{M}} - \mathbf{M}\|_{op} \leq 2\|\lambda\|\|\Phi\|_{op} + d\|\Phi\|_{op}^2 \tag{4.9}
$$

Proof.

$$\|\widehat{\mathbf{M}} - \mathbf{M}\|_{\text{op}} \leq 2\|\sum_{u=1}^{d} \mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{e}_u)\Phi(\mathbf{I}, \mathbf{I}, \mathbf{e}_u)^\top\|_{\text{op}} + \|\sum_{u=1}^{d} \Phi(\mathbf{I}, \mathbf{I}, \mathbf{e}_u)\Phi(\mathbf{I}, \mathbf{I}, \mathbf{e}_u)^\top\|_{\text{op}}$$

Let $\mathbf{E}_1 := \sum_{u=1}^{d} \mathcal{T}(\mathbf{I}, \mathbf{I}, e_u)\Phi(\mathbf{I}, \mathbf{I}, \mathbf{e}_u)^\top$ and $\mathbf{E}_2 := \sum_{u=1}^{d} \Phi(\mathbf{I}, \mathbf{I}, \mathbf{e}_u)\Phi(\mathbf{I}, \mathbf{I}, \mathbf{e}_u)^\top$ respectively. We have:

$$\mathbf{E}_1 = \sum_{r=1}^{R} \lambda_r \mathbf{a}_r \otimes \Phi(\mathbf{I}, \mathbf{b}_r, \mathbf{c}_r)$$

Then $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$\begin{aligned}\mathbf{x}^\top \mathbf{E}_1 \mathbf{y} &= \sum_{r=1}^{R} \lambda_r \mathbf{a}_r^\top \mathbf{x} \Phi(\mathbf{y}, \mathbf{b}_r, \mathbf{c}_r) \\ &\leq (\sum_{r=1}^{R} \lambda_r \mathbf{a}_r^\top \mathbf{x}) \|\Phi\|_{\text{op}} \|\mathbf{y}\| \|\mathbf{b}_r\| \|\mathbf{c}_r\|\end{aligned}$$

Since $\{\mathbf{a}_r\}_{r=1}^{R}$ are orthogonal, $\forall \mathbf{x} \in \mathbb{R}^d, \exists \mathbf{x}' \in \mathbb{R}^R$ such that $x'_r = \mathbf{a}_r^\top \mathbf{x}$ and $\|\mathbf{x}'\| \leq \|\mathbf{x}\|$. Thus

$$\mathbf{x}^\top \mathbf{E}_1 \mathbf{y} \leq \|\Phi\|_{\text{op}} \sum_{r=1}^{R} \lambda_r x'_r \|\mathbf{y}\| \leq \|\Phi\|_{\text{op}} \|\lambda\| \|\mathbf{x}\| \|\mathbf{y}\|$$

For $\mathbf{E}_2$ (which is a symmetric matrix),

$$\mathbf{x}^\top \mathbf{E}_2 \mathbf{x} = \sum_{u=1}^{d} \|\Phi(\mathbf{x}, \mathbf{I}, \mathbf{e}_u)\|^2 \leq d\|\Phi\|_{\text{op}}^2 \|\mathbf{x}\|^2$$

□

That is, $\|\mathbf{E}_1\| \leq \|\Phi\|_{\text{op}}\|\lambda\|$, and $\|\mathbf{E}_2\| \leq d\|\Phi\|_{\text{op}}^2$.

Lemma 4.16 (Perturbation in initialization step for symmetric case)**.** *For symmetric orthogonal tensor, for the matrix generated with trace-based initialization procedure for matrix subspace iteration of the first component, there exists $\{\lambda_r'\}_{r=1}^R$ satisfies the following:*

$$\widehat{\mathbf{M}} = \sum_{r=1}^{R} \lambda_r' \mathbf{a}_r \otimes \mathbf{a}_r + \Phi_M \tag{4.10}$$

and

$$\|\Phi_M\|_{op} \leq \|\lambda\|\|\Phi\|_{op} + d^{3/2}\|\Phi\|_{op}^2. \tag{4.11}$$

Proof. By the linearity of trace and tensor operators, we have the following results:

$$\widehat{\mathbf{M}} = \mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{v}) + \mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{v}_\phi) + \Phi(\mathbf{I}, \mathbf{I}, \mathbf{v}) + \Phi(\mathbf{I}, \mathbf{I}, \mathbf{v}_\phi) \tag{4.12}$$

where

$$(\mathbf{v})_k = \mathsf{trace}(\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{e}_k)) = \sum_{i=1}^{d}\sum_{r=1}^{R} \lambda_r (a_{ir})^2 a_{kr} = \sum_{r=1}^{R} \lambda_r a_{kr}$$

$$(\mathbf{v}_\phi)_k = \mathsf{trace}(\Phi(\mathbf{I}, \mathbf{I}, \mathbf{e}_k))$$

First we notice that $\|\mathbf{v}_\phi\|$ is upper bounded:

$$\|\mathbf{v}_\phi\|^2 = \sum_{k=1}^{d} \mathsf{trace}^2(\Phi(\mathbf{I}, \mathbf{I}, \mathbf{e}_k)) \leq \sum_{k=1}^{d} (d\|\Phi(\mathbf{I}, \mathbf{I}, \mathbf{e}_k)\|)_{\text{op}}^2 \leq d^3\|\Phi\|_{\text{op}}^2 \tag{4.13}$$

Similarly

$$\|\mathbf{v}\|^2 = \sum_{k=1}^{d}(\sum_{r=1}^{R}\lambda_r a_{kr})^2 = \sum_{k=1}^{d}\sum_{\rho,r}^{R}\lambda_\rho\lambda_r a_{kr}a_{k\rho} = \sum_{r,\rho}\lambda_r\lambda_\rho \mathbf{a}_r^\top \mathbf{a}_\rho = \sum_{r=1}^{R}\lambda_r^2$$

Thus the last two operator norm of terms of Eqn (4.12) can be bounded by

$$\|\Phi\|_{\text{op}}(\|\mathbf{v}\| + \|\mathbf{v}_\phi\|) \leq \|\lambda\|\|\Phi\|_{\text{op}} + d^{3/2}\|\Phi\|_{\text{op}}^2$$

The second term of Eqn (4.12) has the following form

$$\mathcal{T}(\mathbf{I}, \mathbf{I}, \mathbf{v}_\phi) = \sum_{r=1}^{R}\lambda_r \mathbf{c}_r^\top \mathbf{v}_\phi \mathbf{a}_r \otimes \mathbf{a}_r$$

Thus $\exists \mathbf{x} \in \mathbb{R}^R : \|\mathbf{x}\| \leq 1$, such that $\lambda_r' = \lambda_r^2 + \lambda_r \mathbf{x}_r \|\mathbf{v}_\phi\|$ □

Lemma 4.17 (Perturbation in convergence step)**.**

$$\|\mathbf{A}_r^\top \Phi_{(1)}(\mathbf{Q}_{\mathbf{C}_r}^{(k)} \odot \mathbf{Q}_{\mathbf{B}_r}^{(k)})\|_{op} \leq \sqrt{r}\|\Phi\|_{op}$$

$$\|(\mathbf{A}_r^c)^\top \Phi_{(1)}(\mathbf{Q}_{\mathbf{C}_r}^{(k)} \odot \mathbf{Q}_{\mathbf{B}_r}^{(k)})\|_{op} \leq \sqrt{r}\|\Phi\|_{op}$$

Proof.

$$(\mathbf{A}_r^\top \Phi_{(1)}(\mathbf{Q}_{\mathbf{C}_r} \odot \mathbf{Q}_{\mathbf{B}_r}))_{ij} = \sum_{(k,z,u)\in[d]^{\times 3}} \Phi_{kzu}(\mathbf{A}_r)_{ki}(\mathbf{Q}_{\mathbf{B}_r})_{zj}(\mathbf{Q}_{\mathbf{C}_r})_{uj}$$

$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^r$ such that $\|\mathbf{x}\|, \|\mathbf{y}\| \leq 1$:

$$\begin{aligned}\mathbf{x}^\top (\mathbf{A}_r^\top \Phi_{(1)}(\mathbf{Q}_{\mathbf{C}_r} \odot \mathbf{Q}_{\mathbf{B}_r}))\mathbf{y} &= \sum_{i,j \in [r]^{\times 2}} x_i y_j \sum_{k,z,u \in [d]^{\times 3}} \Phi_{kzu} (\mathbf{A}_r)_{ki} (\mathbf{Q}_B)_{zj} (\mathbf{Q}_C)_{uj} \\ &= \sum_{j \in [r]} \Phi(\mathbf{A}_r \mathbf{x}, (\mathbf{Q}_{\mathbf{B}_r})_j, (\mathbf{Q}_{\mathbf{C}_r})_j) y_j\end{aligned}$$

By the definition of tensor operator norm, we have that $\forall 1 \leq j \leq r$:

$$\begin{aligned}\Phi(\mathbf{A}_r \mathbf{x}, (\mathbf{Q}_{\mathbf{B}_r})_j, (\mathbf{Q}_{\mathbf{C}_r})_j) &\leq \|\Phi\|_{\text{op}} \|\mathbf{A}_r \mathbf{x}\| \|(\mathbf{Q}_{\mathbf{B}_r})_j\| \|(\mathbf{Q}_{\mathbf{C}_r})_j\| \\ &\leq \|\Phi\|_{\text{op}} \|\mathbf{A}_r\|_{\text{op}} \|\mathbf{x}\| \\ &= \|\Phi\|_{\text{op}} \|\mathbf{x}\|\end{aligned}$$

Thus

$$\begin{aligned}\mathbf{x}^\top (\mathbf{A}_r^\top \Phi_{(1)}(\mathbf{Q}_{\mathbf{C}_r} \odot \mathbf{Q}_{\mathbf{B}_r}))\mathbf{y} &\leq \|\Phi\|_{\text{op}} \|\mathbf{x}\| \sum_{j=1} y_j \\ &\leq \|\mathbf{y}\|_1 \|\Phi\|_{\text{op}} \|\mathbf{x}\| \\ &\leq \sqrt{r} \|\Phi\|_{\text{op}}\end{aligned}$$

The proof for $\mathbf{A}_r^\top \Phi_{(1)}(\mathbf{Q}_{\mathbf{C}_r} \odot \mathbf{Q}_{\mathbf{B}_r})$ is similar.

□

4.7.7 Proof of theorem 4.7

We prove the theorem by examine the success and convergence rate of the initialization stage (lemma 4.18) and the convergence stage (lemma 4.19).

We first provide a few facts that will be used in the proofs.

Fact 1. *The convex combination of scalars is smaller than the largest scalar. That is,* $\forall \alpha \in [0,1]$*:*

$$\alpha x_1 + (1-\alpha)x_2 \leq \max\{x_1, x_2\}$$

Fact 2. *For all* $\theta \in (0,1)$ *and* $A, B \geq 0$*:*

$$\frac{A}{A+\theta B} \leq \frac{1}{1+\theta\frac{B}{A}} \leq \frac{1}{(1+\frac{B}{A})^\theta} = (\frac{A}{A+B})^\theta$$

Lemma 4.18 (Initialization step for noisy tensors)**.** *If the operator norm of the noise tensor is bounded in the following way with a small enough constant* δ_0*:*

$$\|\Phi\|_{op} \leq \min\{\delta_0 \frac{\lambda_r^2 - \lambda_{r+1}^2}{8\|\lambda\|}, \sqrt{\delta_0}\frac{\lambda_r - \lambda_{r+1}}{2\sqrt{d}}\}$$

Then with probability $1 - \mathcal{O}(\delta_0)$ *matrix subspace iteration procedure yields a* r*-sufficient initialization in* $\mathcal{O}(1)$ *time. To be more specific, the tangent value of the subspace angle converges with a rate* $|\frac{\lambda_{r+1}}{\lambda_r}|$*.*

Proof. For matrix subspace iteration of $\widehat{\mathbf{M}} = \mathbf{M} + \Phi_{\mathbf{M}} = \mathbf{A}\mathbf{D}\mathbf{A}^\top + \Phi_{\mathbf{M}}$, we have the

following:

$$
\begin{aligned}
t_{\mathbf{A}_r}^{(k+1)} &\leq \frac{\sigma_{\max}((\mathbf{A}_r^c)^\top \mathbf{A} D \mathbf{A}^\top \mathbf{Q}_{\mathbf{A}_r}) + \sigma_{\max}((\mathbf{A}_r^c)^\top \Phi_{\mathbf{M}} \mathbf{Q}_{\mathbf{A}_r}^{(k)})}{\sigma_{\min}(\mathbf{A}_r^\top \mathbf{A} D \mathbf{A}^\top \mathbf{Q}_{\mathbf{A}_r}) - \sigma_{\max}(\mathbf{A}_r^\top \Phi_{\mathbf{M}} \mathbf{Q}_{\mathbf{A}_r}^{(k)})} \\
&\leq \frac{d_{r+1} \sin\theta_A^k + \|\Phi_{\mathbf{M}}\|_{\text{op}}}{d_r \cos\theta_A^k - \|\Phi_{\mathbf{M}}\|_{\text{op}}}
\end{aligned}
$$

where θ_A^k is the principle angle between the subspace spanned by $\mathbf{A}_r$ and $\mathbf{Q}_{\mathbf{A}_r}^{(k)}$, and $t_{\mathbf{A}_r}^{(k)}$ is $\tan\theta_A^k$.

Let u denote $\frac{\|\Phi_{\mathbf{M}}\|_{\text{op}}}{\mathsf{gap}_r' \cos\theta_A^k}$, where $\mathsf{gap}_r' := d_r - d_{r+1}$. We have:

$$
\begin{aligned}
t_{\mathbf{A}_r}^{(k+1)} &\leq \frac{d_{r+1}\sin\theta_A^k + u\mathsf{gap}_r' \cos\theta_A^k}{d_r \cos\theta_A^k - u\mathsf{gap}_r' \cos\theta_A^k} \\
&\leq \frac{d_{r+1}}{d_r - u\mathsf{gap}_r'} t_{\mathbf{A}_r}^{(k)} + \frac{u\mathsf{gap}_r'}{d_r - u\mathsf{gap}_r'} \\
&= \frac{d_r - 2u\mathsf{gap}_r'}{d_r - u\mathsf{gap}_r'} \cdot \frac{d_{r+1}}{d_r - 2u\mathsf{gap}_r'} t_{\mathbf{A}_r}^{(k)} + \frac{u\mathsf{gap}_r'}{d_r - u\mathsf{gap}_r'} \cdot 1 \\
&\leq \max\{\frac{d_{r+1}}{d_r - 2u\mathsf{gap}_r'} t_{\mathbf{A}_r}^{(k)}, 1\} \qquad \text{(By Fact 1)} \\
&= \max\{\frac{d_{r+1}}{d_{r+1} + (1-2u)\mathsf{gap}_r'} t_{\mathbf{A}_r}^{(k)}, 1\} \\
&\leq \max\{(\frac{d_{r+1}}{d_r})^\theta t_{\mathbf{A}_r}^{(k)}, 1\} \qquad \text{(By Fact 2)}
\end{aligned}
$$

where $\theta := 1 - 2u \leq 1$. Since $\Pr\{\cos\theta_A^0 > 0\} = 1$, by bounding $\|\Phi_{\mathbf{M}}\|_{\text{op}} \leq \frac{d_r - d_{r+1}}{2}\delta_0$ with small enough constant δ_0, combined with Proposition $B.2$ in Wang and Lu [2017], we can verify that $2u \leq 1$ with probability 1 - $\mathcal{O}(\delta_0)$. It is worth noticing that in the noiseless case, we can find a good initialization for matrix subspace iteration with probability 1.

By lemma 4.15, for the slice based initialization, $d_r = \lambda_r^2$, and $\|\Phi_{\mathbf{M}}\| \leq 2\|\lambda\|\|\Phi\|_{\text{op}} + d\|\Phi\|_{\text{op}}^2$, we have $1 - 2u \geq 0$ by bounding:

$$\|\Phi\|_{\text{op}} \leq \min\{\delta_0 \frac{\lambda_r^2 - \lambda_{r+1}^2}{8\|\lambda\|}, \sqrt{\delta_0}\frac{\lambda_r - \lambda_{r+1}}{2\sqrt{d}}\}$$

□

Lemma 4.19 (Convergence step for noisy tensors)**.** *Assume we have the noise tensor bounded in operator norm such that:*

$$\|\Phi\|_{op} \leq \frac{1}{2\sqrt{2}}\frac{\epsilon' \mathsf{gap}_r}{\sqrt{r}} \tag{4.14}$$

where

$$\mathsf{gap}_r := \lambda_r - \lambda_{r+1}$$

Then we have either (1) $t_{\mathbf{A}_r}$ *is small enough:*

$$t_{\mathbf{A}_r}^{(k+1)} \leq \epsilon'$$

Or (2) *converges by the following rule:*

$$t_{\mathbf{A}_r}^{(k+1)} \leq (\frac{\lambda_{r+1}}{\lambda_r})^{\theta} t_{\mathbf{B}_r}^{(k)} t_{\mathbf{C}_r}^{(k)}$$

where

$$\theta := 1 - \frac{2}{\mathsf{gap}_r}(\frac{\sqrt{2}}{\epsilon'} + 1)\sqrt{r}\|\Phi\|_{op} \tag{4.15}$$

Proof. The proof for Theorem 4.19 follows the same style of Lemma $B.1$ in Wang and Lu [2017]. Similar to the noiseless case, we have:

$$t_{\mathbf{A}_r}^{(k+1)} \leq \frac{\lambda_{r+1}\sin\theta_B^{(k)}\sin\theta_C^{(k)} + \sigma_{\max}((\mathbf{A}_r^c)^\top \Phi_{(1)}(\mathbf{Q}_{C_{(r)}}^{(k)} \odot \mathbf{Q}_{\mathbf{B}_{(r)}}^{(k)}))}{\lambda_r \cos\theta_B^{(k)}\cos\theta_C^{(k)} - \sigma_{\max}(\mathbf{A}_r^\top \Phi_{(1)}(\mathbf{Q}_{C_{(r)}}^{(k)} \odot \mathbf{Q}_{\mathbf{B}_{(r)}}^{(k)}))}$$

where θ_U^k is the principle angle between the subspace spanned by $\mathbf{U}_r$ and $\mathbf{Q}_{\mathbf{U}_r}^{(k)}$ for $U \in \{A, B.C\}$, and $t_{\mathbf{A}_r}^{(k)}$ is $\tan\theta_A^k$. Let σ denote the maximum of $\sigma_{\max}((\mathbf{A}_r^c)^\top \Phi_{(1)}(\mathbf{Q}_{C_{(r)}}^{(k)} \odot \mathbf{Q}_{\mathbf{B}_{(r)}}^{(k)}))$ and $\sigma_{\max}(\mathbf{A}_r^\top \Phi_{(1)}(\mathbf{Q}_{C_{(r)}}^{(k)} \odot \mathbf{Q}_{\mathbf{B}_{(r)}}^{(k)}))$, and let $r_1 := \frac{\sqrt{2}\sigma}{\epsilon' \mathsf{gap}_r}$, $r_2 := \frac{2\sigma}{\mathsf{gap}_r}$. Thus

$$\begin{aligned} t_{\mathbf{A}_r}^{(k+1)} &\leq \frac{\lambda_{r+1}\sin\theta_B^{(k)}\sin\theta_C^{(k)} + \sigma}{\lambda_r \cos\theta_B^{(k)}\cos\theta_C^{(k)} - \sigma} \\ &= \frac{\lambda_{r+1}\sin\theta_B^{(k)}\sin\theta_C^{(k)} + r_1 \mathsf{gap}_r \epsilon' \frac{\sqrt{2}}{2}}{\lambda_r \cos\theta_B^{(k)}\cos\theta_C^{(k)} - \frac{1}{2} r_2 \mathsf{gap}_r} \end{aligned}$$

For bounded $\theta_B^{(k)}$ and $\theta_C^{(k)}$ such that $\tan\theta_B^{(k)}$ and $\tan\theta_C^{(k)}$ are less than 1, we have $\cos(\theta_B^{(k)} - \theta_C^{(k)}) \geq \frac{\sqrt{2}}{2}$, and $\cos\theta_B^{(k)}\cos\theta_C^{(k)} \geq \frac{1}{2}$. Thus

$$\begin{aligned} t_{\mathbf{A}_r}^{(k+1)} &\leq \frac{\lambda_{r+1}\sin\theta_B^{(k)}\sin\theta_C^{(k)} + r_1 \mathsf{gap}_r \epsilon' \cos(\theta_B^{(k)} - \theta_C^{(k)})}{\lambda_r \cos\theta_B^{(k)}\cos\theta_C^{(k)} - r_2 \mathsf{gap}_r \cos\theta_B^{(k)}\cos\theta_C^{(k)}} \\ &= \frac{\lambda_{r+1} + r_1 \mathsf{gap}_r \epsilon'}{\lambda_r - r_2 \mathsf{gap}_r} \frac{\sin\theta_B^{(k)}\sin\theta_C^{(k)}}{\cos\theta_B^{(k)}\cos\theta_C^{(k)}} + \frac{r_1 \mathsf{gap}_r}{\lambda_r - r_2 \mathsf{gap}_r}\epsilon' \\ &= \frac{\lambda_{r+1} + r_1 \mathsf{gap}_r \epsilon'}{\lambda_r - r_2 \mathsf{gap}_r} t_{\mathbf{B}_r}^{(k)} t_{\mathbf{C}_r}^{(k)} + \frac{r_1 \mathsf{gap}_r}{\lambda_r - r_2 \mathsf{gap}_r}\epsilon' \\ &= \frac{\lambda_{r+1} + r_1 \mathsf{gap}_r \epsilon'}{\lambda_{r+1} + (1 - r_2)\mathsf{gap}_r} t_{\mathbf{B}_r}^{(k)} t_{\mathbf{C}_r}^{(k)} + \frac{r_1 \mathsf{gap}_r}{\lambda_{r+1} + (1 - r_2)\mathsf{gap}_r}\epsilon' \\ &= (1 - \alpha)\frac{\lambda_{r+1} + r_1 \mathsf{gap}_r \epsilon'}{\lambda_{r+1} + (1 - r_1 - r_2)\mathsf{gap}_r} t_{\mathbf{B}_r}^{(k)} t_{\mathbf{C}_r}^{(k)} + \alpha\epsilon' \end{aligned}$$

where

$$\alpha = \frac{r_1 \mathsf{gap}_r}{\lambda_{r+1} + (1 - r_2)\mathsf{gap}_r}$$

Thus

$$t_{\mathbf{A}_r}^{(k+1)} \leq \max\{\frac{\lambda_{r+1} + r_1 \mathsf{gap}_r \epsilon'}{\lambda_{r+1} + (1 - r_1 - r_2)\mathsf{gap}_r} t_{\mathbf{B}_r}^{(k)} t_{\mathbf{C}_r}^{(k)}, \epsilon'\}$$

Similarly,

$$\frac{\lambda_{r+1} + r_1 \mathsf{gap}_r \epsilon'}{\lambda_{r+1} + (1 - r_1 - r_2)\mathsf{gap}_r} = (1 - \beta)\frac{\lambda_{r+1}}{\lambda_{r+1} + (1 - 2r_1 - r_2)\mathsf{gap}_r} + \beta\epsilon'$$

where

$$\beta = \frac{r_1 \mathsf{gap}_r}{\lambda_{r+1} + (1 - r_1 - r_2)\mathsf{gap}_r}$$

Let θ denote $1 - 2r_1 - r_2$. As long as $\theta > 0$ (that is,),

$$\begin{aligned} t_{\mathbf{A}_r}^{(k+1)} &\leq \max\{\max\{\frac{\lambda_{r+1}}{\lambda_{r+1} + \theta \mathsf{gap}_r}, \epsilon'\} t_{\mathbf{B}_r}^{(k)} t_{\mathbf{C}_r}^{(k)}, \epsilon'\} \\ &\leq \max\{\max\{(\frac{\lambda_{r+1}}{\lambda_r})^\theta, \epsilon'\} t_{\mathbf{B}_r}^{(k)} t_{\mathbf{C}_r}^{(k)}, \epsilon'\} \end{aligned}$$

If $\epsilon' \geq (\frac{\lambda_{r+1}}{\lambda_r})^\theta$ and the $r-$sufficient condition is met,

$$t_{\mathbf{A}_r}^{(1)} \leq \epsilon' t_{\mathbf{B}_r}^{(0)} t_{\mathbf{C}_r}^{(0)} \leq \epsilon'$$

Either the convergence requirement is met after the first iteration, or the procedures converges following:

$$t_{\mathbf{A}_r}^{(k+1)} \leq (\frac{\lambda_{r+1}}{\lambda_r})^{\theta} t_{\mathbf{B}_r}^{(k)} t_{\mathbf{C}_r}^{(k)}$$

Combined with lemma 4.17, the condition $\theta > 0$ is equivalent to:

$$\|\Phi\|_{\text{op}} \leq \frac{\mathsf{gap}_r}{2(\frac{\sqrt{2}}{\epsilon'} + 1)\sqrt{r}} \tag{4.16}$$

Condition (4.16) is satisfied $\forall 1 \leq r \leq R$ as long as:

$$\|\Phi\|_{\text{op}} \leq \frac{1}{2(\frac{\sqrt{2}}{\epsilon'} + 1)} \frac{\min_r \mathsf{gap}_r}{\sqrt{R}}$$

□

Chapter 5: Conclusion and future works

We have presented new convergence rate results for global optimization using radial basis function interpolation and an ε-greedy algorithm which randomly chooses between uniform sampling on the entire domain vs. uniform sampling on a local neighborhood of the current-best solution. This sampling method is simple to implement, but captures the key distinction between local and global search that is present in many other algorithms that are not amenable to theoretical analysis. We find that convergence rates are improved when the size of the local search region is made to shrink over time at a suitable rate, i.e., local search concentrates around the current-best solution over time.

The theory of RBF interpolation relies on a connection between the interpolation error and the distribution of design points on the domain. The latter is measured using the mesh norm, which improves when the design points are more evenly spread out; thus, although it is possible to obtain very similar rates for the case where global search is conducted using non-uniform sampling, it is not possible to *improve* the rates using this analytical technique. To obtain such improvements, it would be necessary to develop new theory that makes a closer connection between the error of RBF interpolation and the shape of the underlying function; we leave this problem for future work.

Our contribution to Gaussian process regression includes a large deviation principle

for the variate vector consists of true function values at two distinct points and their respective estimations. A variant of classic Gartner-Ellis theorem with weaker conditions is estiblished to support our analysis. Later we apply our large deviation principle to obtain the convergence rates of the probability of making two erroneous judgements - reporting one solution as being better than another when in reality the opposite is true, and making large minimum estimation error in an optimization problem.

Essentially, the work aims to precisely characterize the heaviness of the tails of a GP. This is a fundamental property that has rarely been approached in prior work. The results reveal high relevance with the density of experimental points throughout the domain. Aligning with the goal to make the probability of error converge to zero as quickly as possible, this work makes sense of an optimal allocation of experimental effort put across the entire design space.

Lastly, we present a super fast method for tensor decomposition problems which can be met when discovering latent variable models over large datasets. Recovering top r components of asymmetric tensors is often required for many learning scenarios. Existing theory for tensor decompositions guarantee results when the tensor is symmetric. Also, in practice, the tensors are noisy due to finite examples, and also inherently asymmetric. Our efficient algorithm can guarantee recovery of tensor factors for an asymmetric noisy tensor.

There has been a critical factor we have not yet addressed in these chapters - noise. The interpolant may be greatly fooled under the presense of noise, however the fooling effect will be relieved in the setting of Gaussian process regression. The posterior mean will account for the noise and produces a curve or surface that is not an interpolation us-

ing the kernel, but an interpolant using a revised kernel function that is not differentiable at zero. Then the theory framework of using interpolation error analysis to handle optimization error in the setting of Gaussian process cannot work in the noisy problem. Some literature review can be done to explore new mathematical tools that are suitable for this case, after which an extension of the accomplished to adapt for noise may then become possible. Another direction is that, as one may notice, in general optimization problems the error are typically related to mesh norm and in the end we cannot avoid the conclusion that uniformity is the most efficient to reduce error over the domain. Intuitively this is acceptable as that once we don't know anything about the function prior to experiments, we might as well just distribute our budget evenly throughout the domain. Although this is good to accept, one might just be unsatisfied with the uniformity and may wonder in what situations uniformly reducing the mesh norm is not the ultimate suggestion. These can all become ideas for future works.

Bibliography

A. R. Conn, K. Scheinberg, and L. N. Vicente. *Introduction to derivative-free optimization*. SIAM, 2009a.

C. M. Giuliani and E. Camponogara. Derivative-free methods applied to daily production optimization of gas-lifted oil fields. *Computers & Chemical Engineering*, 75:60–64, 2015.

T. Eitrich and B. Lang. Efficient optimization of support vector machine learning parameters for unbalanced datasets. *Journal of Computational and Applied Mathematics*, 196 (2):425–436, 2006.

H. H. Bauschke, W. L. Hare, and W. M. Moursi. A derivative-free comirror algorithm for convex optimization. *Optimization Methods and Software*, 30(4):706–726, 2015.

J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Albert S Berahas, Richard H Byrd, and Jorge Nocedal. Derivative-free optimization of noisy functions via quasi-newton methods. *SIAM Journal on Optimization*, 29(2):965–993, 2019.

A. D. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12:2879–2904, 2011.

J. Calvin, G. Gimbutienė, W. O. Phillips, and A. Zilinskas. On convergence rate of a rectangular partition based global optimization algorithm. *Journal of Global Optimization*, 71(1):165–191, 2018.

A. S. Tikhomirov. On the Markov homogeneous optimization method. *Computational Mathematics and Mathematical Physics*, 46(3):361–375, 2006.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction (2nd edition)*. MIT Press, 2018.

T. Kamishima and S. Akaho. Personalized pricing recommender system: multi-stage epsilon-greedy approach. In *Proceedings of the 2nd International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, pages 57–64, 2011.

V. Raykar and P. Agrawal. Sequential crowdsourced labeling as an epsilon-greedy exploration in a Markov decision process. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pages 832–840, 2014.

T. Back. *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford University Press, 1996.

A. Corana, M. Marchesi, C. Martini, and S. Ridella. Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm. *ACM Transactions on Mathematical Software*, 13(3):262–280, 1987.

X. Hu, Y. Shi, and R. Eberhart. Recent advances in particle swarm. In *Proceedings of the 2004 Congress on Evolutionary Computation*, volume 1, pages 90–97, 2004.

J. F. Schutte and A. A. Groenwold. A study of global optimization using particle swarms. *Journal of Global Optimization*, 31(1):93–108, 2005.

X.-S. Yang. Firefly algorithm, Levy flights and global optimization. In *Research and Development in Intelligent Systems*, volume 26, pages 209–218. 2010.

F. Van den Bergh and A. P. Engelbrecht. A study of particle swarm optimization particle trajectories. *Information Sciences*, 176(8):937–971, 2006.

A. I. F. Vaz and L. N. Vicente. A particle swarm pattern search method for bound constrained global optimization. *Journal of Global Optimization*, 39(2):197–219, 2007.

J. E. Orosz and S. H. Jacobson. Finite-time performance analysis of static simulated annealing algorithms. *Computational Optimization and Applications*, 21(1):21–53, 2002.

V. Torczon. On the convergence of pattern search algorithms. *SIAM Journal on Optimization*, 7(1):1–25, 1997.

R. M. Lewis and V. Torczon. Pattern search algorithms for bound constrained minimization. *SIAM Journal on Optimization*, 9(4):1082–1099, 1999.

R. M. Lewis and V. Torczon. Pattern search methods for linearly constrained minimization. *SIAM Journal on Optimization*, 10(3):917–941, 2000.

T. G. Kolda, R. M. Lewis, and V. Torczon. Optimization by direct search: New perspectives on some classical and modern methods. *SIAM Review*, 45(3):385–482, 2003.

C. Audet and J. E. Dennis. Mesh adaptive direct search algorithms for constrained optimization. *SIAM Journal on Optimization*, 17(1):188–217, 2006.

C. Audet, V. Béchard, and S. Le Digabel. Nonsmooth optimization through mesh adaptive direct search and variable neighborhood search. *Journal of Global Optimization*, 41(2): 299–318, 2008.

C. Audet, G. Savard, and W. Zghal. A mesh adaptive direct search algorithm for multi-objective optimization. *European Journal of Operational Research*, 204(3):545–556, 2010.

M. A. Abramson and C. Audet. Convergence of mesh adaptive direct search to second-order stationary points. *SIAM Journal on Optimization*, 17(2):606–619, 2006.

M. J. D. Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in Optimization and Numerical Analysis*, pages 51–67. Springer, 1994.

A. R. Conn, K. Scheinberg, and P. L. Toint. On the convergence of derivative-free methods for unconstrained optimization. In *Approximation Theory and Optimization: Tributes to M.J.D. Powell*, pages 83–108. 1997.

M. J. D. Powell. UOBYQA: unconstrained optimization by quadratic approximation. *Mathematical Programming*, 92(3):555–582, 2002.

S. Shashaani, F. S. Hashemi, and R. Pasupathy. ASTRO-DF: A class of adaptive sampling trust-region algorithms for derivative-free stochastic optimization. *SIAM Journal on Optimization*, 28(4):3145–3176, 2018.

S. M. Wild, R. G. Regis, and C. A. Shoemaker. ORBIT: Optimization by radial basis function interpolation in trust-regions. *SIAM Journal on Scientific Computing*, 30(6): 3197–3219, 2008.

Z.-J. Shi and J. Guo. A new trust region method with adaptive radius. *Computational Optimization and Applications*, 41(2):225–242, 2008.

S. M. Wild and C. Shoemaker. Global convergence of radial basis function trust region derivative-free algorithms. *SIAM Journal on Optimization*, 21(3):761–781, 2011.

A. R. Conn, K. Scheinberg, and L. N. Vicente. Global convergence of general derivative-free trust-region algorithms to first-and second-order critical points. *SIAM Journal on Optimization*, 20(1):387–415, 2009b.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

M. J. Sasena, P. Papalambros, and P. Goovaerts. Exploration of metamodeling sampling criteria for constrained global optimization. *Engineering Optimization*, 34(3):263–278, 2002.

D. Huang, T. T. Allen, W. I. Notz, and R. A. Miller. Sequential kriging optimization using multiple-fidelity evaluations. *Structural and Multidisciplinary Optimization*, 32 (5):369–382, 2006.

E. Vazquez and J. Bect. Pointwise consistency of the kriging predictor with known mean and covariance functions. In *Advances in Model-Oriented Design and Analysis*, pages 221–228. Springer, 2010a.

A. Zhigljavsky and A. Zilinskas. *Stochastic global optimization*. Springer, 2008.

M. D. Buhmann. *Radial basis functions*. Cambridge University Press, 2003.

Zong-Min Wu and Robert Schaback. Local error estimates for radial basis function interpolation of scattered data. *IMA Journal of Numerical Analysis*, 13(1):13–27, 1993.

H.-M. Gutmann. A radial basis function method for global optimization. *Journal of Global Optimization*, 19(3):201–227, 2001.

R. G. Regis and C. A. Shoemaker. Improved strategies for radial basis function methods for global optimization. *Journal of Global Optimization*, 37(1):113–135, 2007.

K. Holmström. An adaptive radial basis function algorithm (ARBF) for expensive black-box global optimization. *Journal of Global Optimization*, 41(3):447–464, 2008.

R. G. Regis and C. A. Shoemaker. Constrained global optimization of expensive black box functions using radial basis functions. *Journal of Global Optimization*, 31(1):153–171, 2005.

R. G. Regis and C. A. Shoemaker. Parallel stochastic global optimization using radial basis functions. *INFORMS Journal on Computing*, 21(3):411–426, 2009.

R. G. Regis and C. A. Shoemaker. Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization. *Engineering Optimization*, 45(5):529–555, 2013.

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1015–1022, 2010.

S. Janson. Maximal spacings in several dimensions. *The Annals of Probability*, 15(1): 274–280, 1987.

N. M. Temme. Asymptotic estimates of Stirling numbers. *Studies in Applied Mathematics*, 89(3):233–243, 1993.

C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

W. R. Scott, W. B. Powell, and H. P. Simão. Calibrating simulation models using the knowledge gradient with continuous parameters. In B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, editors, *Proceedings of the 2010 Winter Simulation Conference*, pages 1099–1109, 2010.

B. Ankenman, B. L. Nelson, and J. Staum. Stochastic kriging for simulation metamodeling. *Operations Research*, 58(2):371–382, 2010.

S. I. Lee, B. Mortazavi, H. A. Hoffman, D. S. Lu, C. Li, B. H. Paak, J. H. Garst, M. Razaghy, M. Espinal, E. Park, D. C. Lu, and M. Sarrafzadeh. A prediction model for functional outcomes in spinal cord disorder patients using Gaussian process regression. *IEEE Journal of Biomedical and Health Informatics*, 20(1):91–99, 2014.

M. Sheibani and G. Ou. The development of Gaussian process regression for effective regional post-earthquake building damage inference. *Computer-Aided Civil and Infrastructure Engineering*, 36(3):264–288, 2021.

J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian optimization of machine learning algorithms. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 2951–2959, 2012.

P. W. Glynn and S. Juneja. A large deviations perspective on ordinal optimization. In R. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, editors, *Proceedings of the 2004 Winter Simulation Conference*, pages 577–585, 2004.

L. Pronzato and W. G. Müller. Design of computer experiments: space filling and beyond. *Statistics and Computing*, 22(3):681–701, 2012.

V. R. Joseph, E. Gul, and S. Ba. Maximum projection designs for computer experiments. *Biometrika*, 102(2):371–380, 2015.

M. E. Johnson, L. M. Moore, and D. Ylvisaker. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26(2):131–148, 1990.

Subhashis Ghosal and Anindya Roy. Posterior consistency of Gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, 34(5):2413–2429, 2006.

Julien Bect, François Bachoc, and David Ginsbourger. A supermartingale approach to Gaussian process based sequential design of experiments. *Bernoulli*, 25(4A):2883–2919, 2019.

A. L. Teckentrup. Convergence of Gaussian process regression with estimated hyperparameters and applications in Bayesian inverse problems. *SIAM/ASA Journal on Uncertainty Quantification*, 8(4):1310–1337, 2020.

W. Wang, R. Tuo, and C. F. J. Wu. On prediction properties of kriging: Uniform error bounds and robustness. *Journal of the American Statistical Association*, 115(530): 920–930, 2020.

N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.

T. Desautels, A. Krause, and J. W. Burdick. Parallelizing exploration-exploitation tradeoffs in Gaussian process bandit optimization. *Journal of Machine Learning Research*, 15:3873–3923, 2014.

D. Pati, A. Bhattacharya, and G. Cheng. Optimal Bayesian estimation in random covariate design with a rescaled Gaussian process prior. *Journal of Machine Learning Research*, 16:2837–2851, 2015.

S. Vakili, V. Picheny, and N. Durrande. Regret bounds for noise-free Bayesian optimization. *arXiv preprint arXiv:2002.05096*, 2020.

R. J. Adler. On excursion sets, tube formulas and maxima of random fields. *Annals of Applied Probability*, 10(1):1–74, 2000.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications (2nd ed.)*. Springer Berlin Heidelberg, 2009.

J. Zhou and I. O. Ryzhov. A new rate-optimal design for linear regression. Technical report, University of Maryland, 2021.

R. van der Hofstad and H. Honnappa. Large deviations of bivariate Gaussian extrema. *Queueing Systems*, 93(3):333–349, 2019.

M. A. Arcones. Large deviations for M-estimators. *Annals of the Institute of Statistical Mathematics*, 58(1):21–52, 2006.

Emmanuel Vazquez and Julien Bect. Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and inference*, 140(11):3088–3095, 2010b.

Holger Wendland. *Scattered data approximation*. Cambridge University Press, 2004.

Milan Lukić and Jay Beder. Stochastic processes with sample paths in reproducing kernel Hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969, 2001.

J. Li and I. O. Ryzhov. Convergence rates of epsilon-greedy global optimization under radial basis function interpolation. Technical report, University of Maryland, 2021.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014a.

Alastair R Hall. *Generalized method of moments*. Oxford university press, 2005.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

Tamara G Kolda. Symmetric orthogonal tensor decomposition is trivial. *arXiv preprint arXiv:1503.01375*, 2015a.

Anima Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*, 2014b.

Po-An Wang and Chi-Jen Lu. Tensor decomposition via simultaneous power iteration. In *International Conference on Machine Learning*, pages 3665–3673, 2017.

Jerome Brachat, Pierre Comon, Bernard Mourrain, and Elias Tsigaridas. Symmetric tensor decomposition. *Linear Algebra and its Applications*, 433(11-12):1851–1872, 2010.

Tamara G Kolda. Numerical optimization for symmetric tensor decomposition. *Mathematical Programming*, 151(1):225–248, 2015b.

Jiawang Nie. Generating polynomials and symmetric tensor decompositions. *Foundations of Computational Mathematics*, 17(2):423–465, 2017.

Anima Anandkumar, Prateek Jain, Yang Shi, and Uma Naresh Niranjan. Tensor vs. matrix methods: Robust tensor decomposition under block sparse perturbations. In *Artificial Intelligence and Statistics*, pages 268–276, 2016.

Navin Goyal, Santosh Vempala, and Ying Xiao. Fourier PCA and robust tensor decomposition. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 584–593. ACM, 2014.

Vatsal Sharan and Gregory Valiant. Orthogonalized als: A theoretically principled tensor decomposition algorithm for practical use. *arXiv preprint arXiv:1703.01804*, 2017.

Yining Wang and Anima Anandkumar. Online and differentially-private tensor decomposition. In *Advances in Neural Information Processing Systems*, pages 3531–3539, 2016.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

Giorgio Tomasi and Rasmus Bro. A comparison of algorithms for fitting the parafac model. *Computational Statistics & Data Analysis*, 50(7):1700–1734, 2006.

Ignat Domanov and Lieven De Lathauwer. Canonical polyadic decomposition of third-order tensors: Reduction to generalized eigenvalue decomposition. *SIAM Journal on Matrix Analysis and Applications*, 35(2):636–660, 2014.

Florian Roemer and Martin Haardt. A closed-form solution for parallel factor (parafac) analysis. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2365–2368. IEEE, 2008.

Nicolaas Klaas M Faber, Rasmus Bro, and Philip K Hopke. Recent developments in candecomp/parafac algorithms: a critical review. *Chemometrics and Intelligent Laboratory Systems*, 65(1):119–137, 2003.

Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. Computation of the canonical decomposition by means of a simultaneous generalized schur decomposition. *SIAM journal on Matrix Analysis and Applications*, 26(2):295–327, 2004.

Volodymyr Kuleshov, Arun Chaganty, and Percy Liang. Tensor factorization via matrix factorization. In *Artificial Intelligence and Statistics*, pages 507–516, 2015.

Daniel L Pimentel-Alarcón. A simpler approach to low-rank tensor canonical polyadic decomposition. In *Communication, Control, and Computing (Allerton), 2016 54th Annual Allerton Conference on*, pages 474–481. IEEE, 2016.

Peizhen Zhu and Andrew V Knyazev. Angles between subspaces and their tangents. *Journal of Numerical Mathematics*, 21(4):325–340, 2013.

Joseph B Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of " Eckart-Young " decomposition. *Psychometrika*, 35(3):283–319, 1970.

Richard A Harshman. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. *UCLA Working Papers in Phonetics*, 16: 1–84, 1970.

F. Mezzadri. How to generate random matrices from the classical compact groups. *ArXiv Mathematical Physics e-prints*, Sep 2006.

Anima Anandkumar, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Yi-Kai Liu. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 917–925, 2012.

Tiefeng Jiang. How many entries of a typical orthogonal matrix can be approximated by independent normals? *The Annals of Probability*, 34(4):1497–1529, 07 2006. doi: 10.1214/009117906000000205. URL `https://doi.org/10.1214/009117906000000205`.

Shuangzhe Liu and Gõtz Trenkler. Hadamard, khatri-rao, kronecker and other matrix products. *Int. J. Inf. Syst. Sci*, 4(1):160–177, 2008.

Peter Arbenz, Daniel Kressner, and DME Zürich. Lecture notes on solving large scale eigenvalue problems. *D-MATH, EHT Zurich*, 2, 2012.