Moderate deviations inequalities for Gaussian process regression

Jialin Li Ilya O. Ryzhov

October 30, 2022

Abstract

Gaussian process regression is widely used to model an unknown function on a continuous domain by interpolating a discrete set of observed design points. We develop a theoretical framework 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 at an arbitrary point. Our inequalities connect these probabilities to the mesh norm, which measures how well the design points fill the space.

1 Introduction

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

$$f(x) = m(x) + \mathcal{E}(x), \qquad x \in D, \tag{1}$$

where $m : D \to \mathbb{R}$ is a pre-specified "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 & Williams, 2006). This is a Bayesian method: 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 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 & 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 paper 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 moderate 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

$$\pi_n(x, x^*) = P\left(\hat{f}_n(x) \leq \hat{f}_n(x^*) - \delta \,|\, f(x) \geq f(x^*)\right), \tag{2}$$

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 (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

$$\pi_n(x, x^*) \leq C_1 \exp\left(-\delta^2 C_2 h_n^{-\frac{1}{2}s}\right) \tag{3}$$

where $C_1, C_2, s > 0$ 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) as a moderate deviations rate.

The second application deals with the estimation error $\left|\hat{f}_n(x) - f(x)\right|$ at arbitrary $x \in D$. For this error, we prove the uniform bound

$$\sup_{x \in D} P\left(\left|\hat{f}_n(x) - f(x)\right| \geq \delta\right) \leq C_1' \exp\left(-\delta^2 C_2' h_n^{-s}\right). \tag{4}$$

Both types of error probabilities are of broad interest in simulation, statistics, and uncertainty quantification. In particular, the pairwise comparison in (2) is motivated by the approach developed

by Glynn & 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 (2) can only be viewed as an error conditionally given $f(x) \geq f(x^*)$.

Interestingly, the mesh norm is in use as a criterion for design of experiments, in the literature on so-called space-filling designs (Pronzato & 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), we can see that this has the effect of speeding up the rate at which (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 & Roy (2006) or Bect et al. (2019). The rate at which $\hat{f}_n$ converges to f has been studied, e.g., by Teckentrup (2020) and Wang et al. (2020), but these papers do not consider tail probabilities, and so their rates have completely different orders than (3)-(4), though 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). 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). Probabilists have investigated tail probabilities (Adler, 2000; Ghosal & Roy, 2006) and excursion probabilities (Chan & Lai, 2006; Cheng & Xiao, 2016) of Gaussian processes, and some of these results also have the form of moderate deviations laws, but they pertain to generic GPs, rather than the GP regression mechanism. Analogously, moderate deviations laws are available for sums of random variables (Beknazaryan et al., 2019) and random PDE models (Li et al., 2018), but these results do not pertain to GP regression either.

To our knowledge, this paper presents the first moderate deviations results for Gaussian process regression estimators. It is well-known (Dembo & 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 & Ryzhov, 2022), extrema of Gaussian vectors (van der Hofstad & 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 & Zeitouni, 2009), which may be of stand-alone interest.

Section 2 describes the GP regression framework, states the assumptions used throughout the paper, and gives important technical preliminaries. Section 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 6 for readability. Section 4 handles several extensions not covered by the main theorem. Section 5 applies our analysis to derive (3) and (4), and presents several more explicit examples. Section 7 concludes.

2 Gaussian process regression and approximation theory

Before we begin the analysis that leads to (3)-(4), it is necessary to understand the definition, construction, and properties of the GP regression estimator $\hat{f}_n$. The computation of the estimator (a form of Bayesian updating) is described in Section 2.1. Our analysis also makes use of an alternative interpretation, originating from approximation theory, of the GP regression model. We apply this theory to study the asymptotic behavior of the posterior covariance function, which is crucial to the tail probabilities of $\hat{f}_n$. The relevant technical preliminaries are given in Section 2.2.

2.1 Definitions and assumptions

Recalling the model in (1), we assume that the mean function m is Lipschitz continuous, and the Gaussian process $\mathcal{E}$ is specified by

$$\mathbb{E}\left(\mathcal{E}\left(x\right)\right) \quad = \quad 0,$$

$$Cov\left(\mathcal{E}\left(x\right),\mathcal{E}\left(x'\right)\right) \quad = \quad k\left(x,x'\right)$$

for all $x, x' \in D$. In one application, we will 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 $v = (v_1, ..., v_n)$ in $\mathbb{R}^n$, we have

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

Without this assumption, the Gaussian process would be degenerate.

In addition, we assume that there exists a function ϕ on $\mathbb{R}_+$ such that $k\left(x,x'\right) = \phi\left(\|x - x'\|\right)$ for all x, x'. Such a ϕ is called a *radial basis function*. We assume that ϕ is twice differentiable at zero with $\phi''\left(0\right) < 0$. 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 & Bect, 2010). For convenience, we introduce the notation

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

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

Given the design points X_n and observations $f\left(X_n\right)$, the posterior distribution of $f\left(x\right)$, at any arbitrary $x \in D$, is Gaussian with mean

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

and variance

$$P_{X_n}\left(x\right) = k\left(x,x\right) - K\left(x, X_n\right) K\left(X_n, X_n\right)^{-1} K\left(X_n, x\right). \tag{5}$$

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 paper, 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.

2.2 Approximation theory

With the assumptions made in Section 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 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

$$\mathcal{L}_{k,X_n} g(x) = \sum_{m=1}^{n} \alpha_m k(x, x_m), \tag{6}$$

where the coefficients α_m solve the linear system

$$\sum_{m=1}^{n} \alpha_m k\left(x_m, x_{m'}\right) = g\left(x_{m'}\right), \qquad m' = 1, ..., n. \tag{7}$$

In fact, Wu & Schaback (1993) presents a more general form where (6)-(7) include additional polynomial functions, but this will not be necessary for our purposes. It can be shown that $\mathcal{L}_{k,X_n} g\left(x\right) = K\left(x, X_n\right) K\left(X_n, X_n\right)^{-1} g\left(X_n\right)$, similar to the calculations used in GP regression.

Let $\tilde{g}$ be the Fourier transform of g, and suppose that the generalized Fourier transform (as defined in Sec. 8.2 of Wendland, 2004) of the function $x \mapsto \phi\left(\|x\|\right)$ exists and coincides with a continuous function $\tilde{\phi}$ on $\mathbb{R}^d \setminus \{0\}$ satisfying

$$0 < \tilde{\phi}\left(x\right) \leq c_{\tilde{\phi}} \|x\|^{-d-s_\infty} \qquad \text{as } \|x\| \to \infty \tag{8}$$

for suitable constants $c_{\tilde{\phi}}, s_\infty > 0$. In particular, the constant s_∞ will play a significant role in our analysis, and it is worth pointing out that this quantity is explicitly computable for a variety of commonly used kernels. For example, for Gaussian kernels s_∞ can take on any arbitrarily large value (this special case is treated separately in Section 5.3), while for the Matérn kernel, Teckentrup (2020) showed that $s_\infty = 2\sigma$ where σ is the kernel smoothness parameter. Other examples are given in Sec. 8.3 of Wendland (2004).

We now define

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

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 these 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 & 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 & Schaback, 1993 and many other papers in pure interpolation theory, e.g., Li & Ryzhov, 2022). 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 $k\left(\cdot, x\right)$ 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 paper.

Lemma 2.1 (Wu & Schaback, 1993)**.** *Fix $\rho > 0$ and assume that the kernel k satisfies (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 (5) satisfies*

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

Lemma 2.2 (Wu & Schaback, 1993)**.** *Fix g satisfying $c_{g,\phi} < \infty$ and assume that the kernel k satisfies (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 2.1, 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 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 derive the inequality of the form

$$\limsup_{n \to \infty} \frac{1}{a_n} \log \mu_n(E) \leq -\inf_{u \in E} I(u), \tag{9}$$

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 the function I and specify a_n and E in a manner that causes (9) to yield results such as (3) or (4).

Inequalities of the form (9) can be obtained by invoking the Gärtner-Ellis theorem from large deviations theory (Dembo & Zeitouni, 2009). In general, the derivation of the function I consists of two main steps. The first step is to derive the scaled limit

$$\Psi(\gamma) = \limsup_{n \to \infty} \frac{1}{a_n} \Psi_n(a_n \gamma), \tag{10}$$

where $\Psi_n(\gamma) = \log \mathbb{E}_{\mu_n}\left(e^{\langle \gamma, Z_n \rangle}\right)$ denotes the cumulant-generating function of Z_n, and $\langle \cdot, \cdot \rangle$ is the usual L^2 inner product on $\mathbb{R}^p$. The scaling Ψ is allowed to take values on the extended number

line (i.e., may be $+\infty$ for some γ). The second step obtains I via the Fenchel-Legendre transform

$$I(u) = \sup_{\gamma \in \mathbb{R}^p} \langle \gamma, u \rangle - \Psi(\gamma) \tag{11}$$

of Ψ. Inequality (9) then follows under certain technical conditions on Ψ.

Our analysis in this section follows this outline. Section 3.1 studies the cumulant-generating functions of $\{Z_n\}$ and characterizes Ψ. Section 3.2 then studies the Fenchel-Legendre transform of Ψ, and Section 3.3 analyzes the convergence rates of various terms in the Fenchel-Legendre transform, which helps us to select $\{a_n\}$ in a way that makes (9) yield an informative inequality (stated in Section 3.4). In Section 5, we will instantiate this inequality with certain choices of E that yield (3)-(4).

One complication that arises in this procedure is that the technical conditions of the classical Gärtner-Ellis theorem do not hold in our setting, so additional analysis is required to obtain (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 6 at the end of the paper; the core result of that section is Theorem 6.1, which recovers the desired inequality. Here, we take (9) as given, referring readers to Theorem 6.1 for the proof, and focus on applying this inequality to the specific sequence $\{Z_n\}$.

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 $A_n m\left(\bar{X}_n\right)$ and covariance matrix $V_n = A_n \Sigma_n A_n^\top$, where

$$A_n = \begin{bmatrix} K(x, X_n) K(X_n, X_n)^{-1} & 0 & 0 \\ K(x^*, X_n) K(X_n, X_n)^{-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \Sigma_n = \begin{bmatrix} K(X_n, X_n) & K(X_n, x) & K(X_n, x^*) \\ K(x, X_n) & k(x, x) & k(x, x^*) \\ k(x^*, X_n) & k(x^*, x) & k(x^*, x^*) \end{bmatrix}.$$

For convenience, we introduce the notation

$$Q_{X_n}(x) \;=\; K(x, X_n) K(X_n, X_n)^{-1} K(X_n, x),$$

$$Q_{X_n}(x, x^*) \quad = \quad K(x, X_n) K(X_n, X_n)^{-1} K(X_n, x^*).$$

Then, the power function P_{X_n} in (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

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

for any $\gamma \in \mathbb{R}^4$. Then, by (10),

$$\Psi(\gamma) = \gamma^\top \left(\lim_{n\to\infty} A_n m\left(\bar{X}_n\right)\right) + \frac{1}{2} \limsup_{n\to\infty} \gamma^\top (a_n V_n) \gamma, \tag{12}$$

provided that the limit on the right-hand side of (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 2.2.

First, we consider the limit of

$$A_n m\left(\bar{X}_n\right) = \left[\; K(x, X_n) K(X_n, X_n)^{-1} m(X_n) \quad K(x^*, X_n) K(X_n, X_n)^{-1} m(X_n) \quad m(x) \quad m(x^*) \; \right].$$

We may observe that $\mathcal{L}_{k,X_n} m(x)$, with $\mathcal{L}_{k,X_n}$ being the operator that maps a function to its interpolant given k and X_n as in (6), 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 whose zeros become dense (as $n \to \infty$) around x and x^*. That is, there exist $\rho, \rho^* > 0$ such that $h_n(B_{x,\rho} \cap D) \to 0$ and $h_n(B_{x^*,\rho^*} \cap D) \to 0$ as $n \to \infty$. Consequently, $m(x) - \mathcal{L}_{k,X_n} m(x) \to 0$, whence $\lim_{n\to\infty} A_n m\left(\bar{X}_n\right) = m_0$, with $m_0 = (m(x), m(x^*), m(x), m(x^*))^\top$. Thus, (12) becomes

$$\Psi(\gamma) = \gamma^\top m_0 + \frac{1}{2} \limsup_{n\to\infty} \gamma^\top (a_n V_n) \gamma. \tag{13}$$

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

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 2.1 we can see that $P_{X_n}(y) \to 0$ if X_n is dense in D as $n \to \infty$. 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 2.2, 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 X_n is dense in D as $n \to \infty$, 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 $\lambda_1 = 2(k(x,x) + k(x,x^*))$, $\lambda_2 = 2(k(x,x) - k(x,x^*))$ 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

$$U_3 = \frac{1}{\sqrt{2}}(1,0,-1,0)^\top, \qquad U_4 = \frac{1}{\sqrt{2}}(0,1,0,-1)^\top. \tag{14}$$

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

$$(U_3', U_4') = (U_3, U_4)\, T \tag{15}$$

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

Looking back to (11) and (13), we can write the Fenchel-Legendre transform as

$$I(u) \;=\; \sup_{\gamma \in \mathbb{R}^4} (u - m_0)^\top \gamma - \frac{1}{2} \limsup_{n\to\infty} \gamma^\top (a_n V_n)\, \gamma$$

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

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

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

as long as $\gamma_j \neq 0$ for $j \in \{1,2\}$. Therefore, the supremum in (11) can only be achieved at γ for which $\gamma_1 = \gamma_2 = 0$, whence

$$
\begin{aligned}
I(u) &= \sup_{\gamma_3,\gamma_4} (u-m_0)^\top U_3\gamma_3 + (u-m_0)^\top U_4\gamma_4 - \frac{1}{2}\limsup_{n\to\infty} \sum_j \left(\sum_i \gamma_i U_i^\top U_{j,n}\right)^2 a_n\lambda_{j,n} \\
&\geq \sup_{\gamma_3,\gamma_4} \left[(u-m_0)^\top U_3\gamma_3 + (u-m_0)^\top U_4\gamma_4 - \frac{1}{2}\sum_{j=1}^{2} \lambda_j \limsup_{n\to\infty} \left(\sum_{i=3}^{4} \gamma_i U_i^\top U_{j,n}\right)^2 a_n \right. \\
&\quad \left. -\frac{1}{2}\sum_{j=3}^{4} \limsup_{n\to\infty} \left(\sum_{i=3}^{4} \gamma_i U_i^\top U_{j,n}\right)^2 \limsup_{n\to\infty} a_n\lambda_{j,n} \right] \\
&= \sup_{\gamma_3,\gamma_4} \left[(u-m_0)^\top U_3\gamma_3 + (u-m_0)^\top U_4\gamma_4 - \frac{1}{2}\sum_{j=1}^{2} \lambda_j \limsup_{n\to\infty} \left(\sum_{i=3}^{4} \gamma_i U_i^\top U_{j,n}\right)^2 a_n \right. \\
&\quad \left. -\frac{1}{2}\sum_{j=3}^{4} \left(T_{1,j-2}\gamma_3 + T_{2,j-2}\gamma_4\right)^2 \limsup_{n\to\infty} a_n\lambda_{j,n} \right]. \qquad (16)
\end{aligned}
$$

The supremum value in (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 (16) finite. Some matrix perturbation analysis is required to understand the rate that a_n can take.

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

$$\nu_{ijn} \lambda_{j,n} = \nu_{ijn} \lambda_i - U_i^\top \tilde{V} U_{j,n}. \tag{17}$$

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

$$\nu_{ijn} = -\frac{U_i^\top \tilde{V} U_{j,n}}{\lambda_{j,n}}, \quad i \in \{3,4\}, \, j \in \{1,2\}. \tag{18}$$

This allows us to bound the limits superior in (16) as shown in Lemmas 3.1 and 3.2 below.

Lemma 3.1. *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 (18), 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 (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 3.2. *Suppose that the matrix T defined in (15) has no zero-valued entries. Then,*

$$\lambda_{j,n} = \left(\frac{1}{2} + o(1)\right)\left(P_{X_n}(x) + \frac{T_{2,j-2}}{T_{1,j-2}} P_{X_n}(x, x^*)\right) = \left(\frac{1}{2} + o(1)\right)\left(P_{X_n}(x^*) + \frac{T_{1,j-2}}{T_{2,j-2}} P_{X_n}(x, x^*)\right). \tag{19}$$

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

Proof: Recall (17) 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, (18) can be rewritten as

$$\lambda_{j,n} = -\frac{U_i^\top \tilde{V} U_{j,n}}{\nu_{ijn}}, \quad i, j \in \{3, 4\}. \tag{20}$$

The first equality in (19) can be obtained by setting $i = 3$, whence (20) 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

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

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 (19) can be obtained by repeating the above arguments with $i = 4$. □

The analysis in Lemma 3.2 is easily extended to handle situations where T has zero-valued entries. If this occurs, we must have $||T_{11}| - |T_{21}|| = 1$ because T is orthonormal. In the first case, we can repeat the proof of Lemma 3.2 with $i = 4, j = 3$ and $i = 3, j = 4$ and obtain

$$\lambda_{3,n} = \left(\frac{1}{2} + o(1)\right) P_{X_n}(x^*), \qquad \lambda_{4,n} = \left(\frac{1}{2} + o(1)\right) P_{X_n}(x). \tag{21}$$

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 3.1 and 3.2 depend on $P_{X_n}(x, x^*)$, which is a difficult object to study. Lemma 3.3 establishes a bound that relates this quantity to a simpler function of the design points. Then, Lemma 3.4 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 (16).

Lemma 3.3. *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

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

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

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

which completes the proof. □

Lemma 3.4. *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 3.3 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 3.3, 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 $\lambda_{\min}(K(X_n', X_n'))$ or $\lambda_{\min}\left(K\left(\bar{X}_n, \bar{X}_n\right)\right)$ converges to zero. For this, we cite the following result (Theorem 12.3 of Wendland, 2004).

Lemma 3.5 (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,*

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

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

Combining Lemmas 3.3-3.5, 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 3.3 becomes

$$2P_{X_n}(x, x^*) + P_{X_n}(x) + P_{X_n}(x^*) \geq 2C_d \phi_0\left(\frac{M_d}{q_{\bar{X}_n}}\right) q_{\bar{X}_n}^{-d}. \tag{22}$$

The lower bound in (22) can be connected back to the upper bound obtained in Lemma 3.2 in the following manner. Note that, if T has no zero-valued entries as assumed in Lemma 3.2, 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 (22) to (19) with $j = 4$ and argue

$$\begin{aligned}
\lambda_{4,n} &\leq \left(\frac{1}{2} + o(1)\right)\left[P_{X_n}(x) - \frac{T_{22}}{2T_{12}}\left(P_{X_n}(x) + P_{X_n}(x^*) - 2C_d\phi_0\left(\frac{M_d}{q_{\bar{X}_n}}\right)q_{\bar{X}_n}^{-d}\right)\right] && (23)\\
&= \left(\frac{1}{2} + o(1)\right)\left[\left(1 - \frac{T_{22}}{2T_{12}}\right)P_{X_n}(x) - \frac{T_{22}}{2T_{12}}P_{X_n}(x^*) + O\left(q_{\bar{X}_n}^{s_\infty}\right)\right], && (24)\\
&= O\left(h_n\left(B_{x,\rho} \cap D\right)^{s_\infty}\right). && (25)
\end{aligned}$$

In this derivation, (23) uses the fact that $\frac{T_{12}}{T_{22}} < 0$ to convert the lower bound in (22) into an upper bound, while (24) applies (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 (25) by applying Lemma 2.1 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 (19) 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 2.1, 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 2.1 to the bound in Lemma 3.1 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 X_n is dense in D as $n \to \infty$, we have $h_n\left(B_{x,\rho} \cap D\right) \leq h_n(D)$ with $h_n(D) \to 0$. Thus, among the limits superior in (16), 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 $\lambda_{4,n}$.

3.4 Main moderate deviations inequality

The conclusions of Section 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 (16), and bound $I(u) \geq I^l(u)$, where I^l is defined as

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

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

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

when $\frac{T_{12}}{T_{22}} > 0$, for some suitable constants c_3, c_4. Furthermore, recalling (14) and the definition of m_0, we find that $m_0^\top U_3 = m_0^\top U_4 = 0$. Now, applying (9), we can finally state our main result.

Theorem 3.1. *Let T be as in (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 $||T_{11}| - |T_{21}|| \notin \{0,1\}$, we have*

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(E) \leq -\inf_{u\in E} I^l(u) \tag{26}$$

for any closed $E \subseteq \mathbb{R}^4$, with

$$I^l(u) = \sup_{\gamma_3,\gamma_4 \in \mathbb{R}} u^\top U_3\gamma_3 + u^\top U_4\gamma_4 - \frac{1}{2}c_l (|T_{11}|\gamma_3 + |T_{21}|\gamma_4)^2. \tag{27}$$

Throughout this analysis, we have assumed that X_n is dense in D when $n \to \infty$, but it is possible to recover Theorem 3.1, 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 $\min\left\{h_n (B_{x,\rho})^{-\frac{1}{2}s_\infty}, h_n (B_{x^*,\rho})^{-\frac{1}{2}s_\infty}\right\}$.

The right-hand side of (26) is some strictly negative, problem-specific constant, and it is the order of a_n that governs the convergence rate of $\mu_n(E)$. The moderate deviations inequality allows the rate to depend on the kernel through the quantity s_∞, but otherwise the complex interdependence of the various elements of $K(X_n, X_n)$ has been streamlined using Lemma 2.1 and Lemma 3.5. For this reason, the bound in (26) may not be the tightest possible.

4 Extensions and special cases

In this section, we treat several special cases not covered by Theorem 3.1. First, Section 4.1 handles the situation where $||T_{11}| - |T_{21}|| \in \{0,1\}$. Section 4.2 discusses how the bound of Theorem 3.1 is improved if we only consider one reference point x instead of two points x, x^*.

4.1 Special cases of Theorem 3.1

Again, let T be as in (15), and suppose that $||T_{11}| - |T_{21}|| = 0$. For simplicity, we consider the case $T_{11} = T_{21}$ as the other cases are very similar.

Because T is orthonormal, we must have $T_{11} = T_{21} = \frac{\sqrt{2}}{2}$. Then, returning to the derivation of (16) and recalling that $m_0^\top U_3 = m_0^\top U_4 = 0$, we have

$$\begin{aligned} I(u) &\geq \sup_{\gamma_3,\gamma_4} u^\top U_3\gamma_3 + u^\top U_4\gamma_4 - \frac{1}{2}\limsup_{n\to\infty}\left[\sum_{j=1}^{2}\lambda_j\left(\sum_{i=3}^{4}\gamma_i U_i^\top U_{j,n}\right)^2 a_n \right. \\ &\quad \left. -\frac{1}{2}\left((\gamma_3+\gamma_4)^2\lambda_{3,n} + (\gamma_3-\gamma_4)^2\lambda_{4,n}\right)a_n\right] \end{aligned} \tag{28}$$

Using a similar argument as in Lemma 3.1, we explicitly derive

$$\begin{aligned} \sum_{j=1}^{2}\lambda_j\left(\sum_{i=3}^{4}\gamma_i U_i^\top U_{j,n}\right)^2 &= \frac{1+o(1)}{8}\left[\frac{1}{\lambda_1}\left(\gamma_3 P_{X_n}(x) + (\gamma_3+\gamma_4)P_{X_n}(x,x^*) + \gamma_4 P_{X_n}(x^*)\right)^2 \right. \\ &\quad \left. +\frac{1}{\lambda_2}\left(\gamma_3 P_{X_n}(x) + (\gamma_4-\gamma_3)P_{X_n}(x,x^*) - \gamma_4 P_{X_n}(x^*)\right)^2\right]. \end{aligned} \tag{29}$$

By direct application of Lemma 3.2, we have

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

Let us now take $a_n \sim h_n^{-s_\infty}$, which guarantees $\limsup_{n\to\infty} a_n P_{X_n}(x) \leq c_l$ and $\limsup_{n\to\infty} a_n P_{X_n}(x^*) \leq c_l$ for some finite $c_l > 0$. This allows us to place a bound on (28) that drops all negligible terms in (29) that contain $P_{X_n}(x)$ and $P_{X_n}(x^*)$. That is,

$$\begin{aligned} I(u) &\geq \sup_{\gamma_3,\gamma_4} u^\top U_3\gamma_3 + u^\top U_4\gamma_4 - \frac{1}{16}\limsup_{n\to\infty}\left[\frac{1}{\lambda_1}(\gamma_3+\gamma_4)^2 P_{X_n}^2(x,x^*) + \frac{1}{\lambda_2}(\gamma_4-\gamma_3)^2 P_{X_n}^2(x,x^*) \right. \\ &\quad \left. +2(\gamma_3+\gamma_4)^2 P_{X_n}(x) + 2(\gamma_3-\gamma_4)^2 P_{X_n}(x^*) - 4\gamma_3\gamma_4 P_{X_n}(x,x^*)\right]a_n \end{aligned} \tag{30}$$

The last term in (30) requires us to take $\gamma_3\gamma_4 \leq 0$ to avoid $I(u)$ becoming $-\infty$. Then, (30) becomes

$$I(u) \geq \sup_{\gamma_3\gamma_4\leq 0} u^\top U_3\gamma_3 + u^\top U_4\gamma_4 - \frac{1}{16}\limsup_{n\to\infty}\left[\frac{1}{\lambda_1}(\gamma_3+\gamma_4)^2 P_{X_n}^2(x,x^*) + \frac{1}{\lambda_2}(\gamma_4-\gamma_3)^2 P_{X_n}^2(x,x^*)\right.$$

$$+2\left(\gamma_3+\gamma_4\right)^2 P_{X_n}\left(x\right)+2\left(\gamma_3-\gamma_4\right)^2 P_{X_n}\left(x^*\right)-4\gamma_3\gamma_4\left(P_{X_n}\left(x\right)+P_{X_n}\left(x^*\right)\right)\Big] a_n \tag{31}$$

$$\geq \sup_{\gamma_3\gamma_4\leq 0} u^\top U_3\gamma_3+u^\top U_4\gamma_4-\frac{c_l}{16}\left[\frac{1}{\lambda_1}\left(\gamma_3+\gamma_4\right)^2 c^2_{k(\cdot,x^*),\phi}+\frac{1}{\lambda_2}\left(\gamma_3-\gamma_4\right)^2 c^2_{k(\cdot,x^*),\phi}\right.$$
$$\left.+2\left(\left(\gamma_3+\gamma_4\right)^2+\left(\gamma_3-\gamma_4\right)^2-4\gamma_3\gamma_4\right)\right] \tag{32}$$

$$= \sup_{\gamma_3\gamma_4\leq 0} u^\top U_3\gamma_3+u^\top U_4\gamma_4-\frac{c_l}{16}\left[\frac{1}{\lambda_1}\left(\gamma_3+\gamma_4\right)^2 c^2_{k(\cdot,x^*),\phi}+\left(\gamma_3-\gamma_4\right)^2\left(\frac{c^2_{k(\cdot,x^*),\phi}}{\lambda_2}+4\right)\right].$$

where (31) applies Lemma 3.3 to bound $P_{X_n}\left(x,x^*\right)$, and (32) uses the bound $P^2_{X_n}\left(x,x^*\right)\leq c^2_{k(\cdot,x^*),\phi}P_{X_n}\left(x\right)$ from Lemma 2.2. After some algebra, we obtain a version of the moderate deviations inequality (26) with

$$I^l\left(u\right) = \sup_{\gamma_3\gamma_4\leq 0}\frac{1}{2}u^\top\left(U_3+U_4\right)\left(\gamma_3+\gamma_4\right)+\frac{1}{2}u^\top\left(U_3-U_4\right)\left(\gamma_3-\gamma_4\right)$$
$$-\frac{c_l}{16}\left[\frac{1}{\lambda_1}\left(\gamma_3+\gamma_4\right)^2 c^2_{k(\cdot,x^*),\phi}+\left(\gamma_3-\gamma_4\right)^2\left(\frac{c^2_{k(\cdot,x^*),\phi}}{\lambda_2}+4\right)\right]. \tag{33}$$

The second special case $||T_{11}|-|T_{21}||=1$ is handled in almost the same way. For simplicity, we take $T_{11}=T_{22}=0$ as the other possible cases are very similar. Then, (29) remains unchanged, but we use (21) instead of Lemma 3.2. We can thus omit the final cross term $-4\gamma_3\gamma_4P_{X_n}\left(x,x^*\right)$ in (30), so we no longer need $\gamma_3\gamma_4\leq 0$. The other arguments are unchanged and yield another version of (26) with

$$I^l\left(u\right) = \sup_{\gamma_3,\gamma_4}\frac{1}{2}u^\top\left(U_3+U_4\right)\left(\gamma_3+\gamma_4\right)+\frac{1}{2}u^\top\left(U_3-U_4\right)\left(\gamma_3-\gamma_4\right)$$
$$-\frac{c_l}{16}\left[\left(\gamma_3+\gamma_4\right)^2\left(\frac{c^2_{k(\cdot,x^*),\phi}}{\lambda_1}+2\right)+\left(\gamma_3-\gamma_4\right)^2\left(\frac{c^2_{k(\cdot,x^*),\phi}}{\lambda_2}+2\right)\right]. \tag{34}$$

4.2 Moderate deviations for a single point

In the following, we treat the special case where $Z_n=\left(\hat{f}_n\left(x\right),f\left(x\right)\right)^\top$, for a single, fixed $x\in D$. Though one can obtain moderate deviations inequalities for such Z_n directly from Theorem 3.1, it is possible to tighten the bounds by modifying the analysis. Below we highlight those parts of the argument that require changes.

As before, we let μ_n denote the probability law of Z_n and assume that the set X_n of design

points becomes dense in D as $n \to \infty$. We write $Z_n = A_n f(X_n')$, where $X_n' = X_n \cup \{x\}$ and

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

The covariance matrix of Z_n is given by

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

where Q_{X_n} is as defined in Section 3.1. Equation (13) remains unchanged with $m_0 = (m(x), m(x))^\top$. Repeating the arguments in Section 3.2, we have $V_n \to V$ entrywise, where V is a 2×2 matrix whose elements are all equal to $k(x,x)$. This matrix has two eigenvalues $\lambda_1 > 0$ and $\lambda_2 = 0$, with corresponding eigenvectors

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

Similarly, denote by $\lambda_{i,n}$ and $U_{i,n}$ the eigenvalues and eigenvectors of V_n.

Again following Section 3.2, the supremum in (11) can only be achieved at $\gamma \in \mathbb{R}^2$ satisfying $\gamma_1 = 0$. Repeating the derivation of (16), we obtain

$$I(u) \geq \sup_{\gamma_2} (u - m_0)^\top U_2 \gamma_2 - \frac{1}{2}\lambda_1 \left(\gamma_2 U_2^\top U_{1,n}\right)^2 a_n - \frac{1}{2} \limsup_{n\to\infty} \left(\gamma_2 U_2^\top U_{2,n}\right)^2 \limsup_{n\to\infty} a_n \lambda_{2,n}$$

where $\limsup_{n\to\infty} \left(\gamma_2 U_2^\top U_{2,n}\right)^2 = \gamma_2^2$ and

$$\left(\gamma_2 U_2^\top U_{1,n}\right)^2 = \frac{1}{\lambda_{1,n}^2} \left(\gamma_2 U_{1,n}^\top (V - V_n) U_2\right)^2 \leq \frac{\gamma_2^2}{\lambda_{1,n}^2} P_{X_n}^2(x) = O\left(P_{X_n}^2(x)\right),$$

analogously to Lemma 3.1. Repeating the proof of Lemma 3.2, we obtain

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

The rest of Section 3.3 remains the same, but it is sufficient to use Lemma 3.4, as there is no longer any cross term. Thus, we avoid the factor of 2 in the lower bound derived in Lemma 3.3, which allows us to take $a_n \sim h_n^{-s_\infty}$. It then follows that $\left(\gamma_2 U_2^\top U_{1,n}\right)^2 a_n \to 0$, while $\limsup_{n\to\infty} a_n \lambda_{2,n} \leq c_l$ for some $c_l < \infty$. We thus obtain $I(u) \geq I^l(u)$, where

$$\begin{aligned} I^l(u) &= \sup_{\gamma_2} (u - m_0)^\top U_2 \gamma_2 - \frac{1}{2} c_l \gamma_2^2 \\ &= \sup_{\gamma_2} u^\top U_2 \gamma_2 - \frac{1}{2} c_l \gamma_2^2 \\ &= \frac{(u_1 - u_2)^2}{4 c_l}. \end{aligned}$$

We now apply (9) to obtain the desired result. Note that the support set $\left\{\gamma \in \mathbb{R}^2 : \Psi(\gamma) < \infty\right\}$ is a subspace, so (9) still holds by the results of Section 6.

Theorem 4.1. *Take* $a_n \sim h_n^{-s_\infty}$ *and let* c_l *be a constant satisfying* $\limsup_{n\to\infty} a_n \lambda_{2,n} \leq c_l$*. Then,*

$$\limsup_{n\to\infty} \frac{1}{a_n} \log \mu_n(E) \leq -\inf_{u\in E} \frac{(u_1 - u_2)^2}{4c_l}.$$

for any closed $E \subseteq \mathbb{R}^2$*.*

By considering one reference point x instead of two points x, x^*, we obtain a better bound because it is no longer necessary to bound a cross term, as in Lemma 3.3. However, the bound will not get worse if we consider $K > 2$ points, i.e., with $Z_n = \left\{\hat{f}_n(x_k^*), f(x_k^*)\right\}_{k=1}^K$ for some $K < \infty$. This is because the same bound in Lemma 3.3 can be applied to every possible cross term.

5 Applications: pairwise comparisons and estimation error

Sections 5.1-5.2 apply Theorems 3.1 and 4.1 to prove (3) and (4), respectively. The proofs are very similar, but use different definitions of the error set E in (26). Section 5.3 presents several other results of interest where the moderate deviations bound can be made more explicit.

5.1 Moderate deviations for false ordering

We return to (2) and write

$$\pi_n(x, x^*) = \frac{P\left(\hat{f}_n(x) \leq \hat{f}_n(x^*) - \delta,\, f(x) \geq f(x^*)\right)}{P(f(x) \geq f(x^*))}. \tag{35}$$

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 with

$$E = \left\{u \in \mathbb{R}^4 : u_1 \leq u_2 - \delta,\, u_3 \geq u_4\right\}. \tag{36}$$

We will apply Theorem 3.1 and derive a more explicit form for (27). First, note that the supremum in (27) can only be finite when

$$\frac{u^\top U_3}{|T_{11}|} = \frac{u^\top U_4}{|T_{21}|}. \tag{37}$$

Letting η be the value in (37), we then have $I^l(u) = \frac{\eta^2}{2c_l}$. Then, we minimize $I^l(u)$ subject to (36)-(37). From the optimality conditions, it can be seen that the inequalities in (36) 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 3.1, we complete the proof of (3). The formal statement of the result is as follows.

Theorem 5.1. *Let T be as in (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 $||T_{11}| - |T_{21}|| \notin \{0,1\}$, 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 (35) 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

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

using the assumption made in Section 2.1 that ϕ is twice differentiable at zero with $\phi''(0) < 0$. Because D is compact, there exists some $c_D > 0$ satisfying

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

Furthermore, the constant C_2 in Theorem 5.1 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 5.1. *Suppose that we are in the situation of Theorem 5.1. Then,*

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

where C_1', C_2, c_D are positive constants.

Finally, we consider two special cases covered in Section 4. First, in the case where $T_{11} = T_{21}$ or $T_{12} = T_{22}$, we apply (33). It can be shown that the supremum is attained when $\gamma_3 + \gamma_4 = 0$, which satisfies the condition $\gamma_3\gamma_4 \leq 0$ and achieves $u^\top(U_3 + U_4) = 0$ and $\left|u^\top U_3 - u^\top U_4\right| = \frac{\delta}{2\sqrt{2}}$. Consequently, we have

$$\pi_n(x,x^*) \leq C_1 \exp\left(-\delta^2 C_2' h_n^{-s_\infty}\right), \tag{38}$$

where C_2' depends on $\lambda_2, c_l, c_{k(\cdot,x^*),\phi}$. In the second special case where $||T_{11}| - |T_{21}|| = 1$, we apply (34). Again, the supremum achieves $u^\top(U_3 + U_4) = 0$ and $\left|u^\top U_3 - u^\top U_4\right| = \frac{\delta}{2\sqrt{2}}$ and (38) follows. In summary, both special cases admit an exponential bound in $h_n^{-s_\infty}$ rather than $h_n^{-\frac{1}{2}s_\infty}$.

5.2 Moderate deviations for pointwise estimation error

The convergence rate of the tail probability $P\left(\left|\hat{f}_n(x) - f(x)\right| \geq \delta\right)$ for fixed x can be obtained from Theorem 4.1 using the error event $E' = \{|u_1 - u_2| \geq \delta\}$. It is easy to see that

$$\inf_{u\in E'} \frac{(u_1-u_2)^2}{4c_l} = \frac{\delta^2}{4c_l},$$

whence we have

$$P\left(\left|\hat{f}_n(x) - f(x)\right| \geq \delta\right) \leq C_1 \exp\left(-\frac{\delta^2 C_2}{4c_l} h_n^{-s_\infty}\right)$$

where $C_1, C_2 > 0$ are constants. As in Corollary 5.1, we can take the supremum over all $x \in D$ to obtain the desired result, which is formally stated as follows.

Theorem 5.2. *There exist constants $C_1, C_2 > 0$ such that*

$$\sup_{x\in D} P\left(\left|\hat{f}_n(x) - f(x)\right| \geq \delta\right) \leq C_1 \exp\left(-\frac{\delta^2 C_2}{4c_l} h_n^{-s_\infty}\right).$$

A natural question is whether this result can be extended to bound $P\left(\sup_x \left|\hat{f}_n(x) - f(x)\right| \geq \delta\right)$, the tail probability of the estimation error over the domain, perhaps by combining Theorem 5.2 with the continuity of f and an epsilon-net partition of D. Such an argument could work if f and

$\hat{f}_n$ were uniformly equicontinuous over all sample paths, a property known as "uniform modulus of continuity" (Marcus & Shepp, 1972). It is possible to obtain such results for generic GPs in certain settings (Ciesielski, 1961), but they are currently not available for the specific mechanism of Gaussian process regression; recent work, such as Lederer et al. (2019) and Toth & Oberhauser (2020), has only bounded the modulus of continuity on a restricted set of sample paths, not almost surely. This extension is an interesting topic for future work.

5.3 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 5.1 in our presentation, but analogs of the other results in Sections 5.1-5.2 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, $k(x, x^*) = \exp\left(-\alpha\|x - x^*\|_2^2\right)$. 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 3.5 for this setting, namely,

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

Thus, using the above results instead of Lemmas 2.1 and 3.5, 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 5.1.

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 5.1.

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 5.1. 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 5.1.

6 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 (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 (11). The classical Gärtner-Ellis theorem (Dembo & Zeitouni, 2009) establishes the inequality (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 (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 (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 6.1), 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 6.1) then proves the inequality for E.

Lemma 6.1. *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: 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

$$\begin{aligned}
\frac{1}{a_n}\log\mu_n^{\mathcal{P}}(\mathcal{P}B_{u,\rho^u}) &\leq \frac{1}{a_n}\log\mathbb{E}\left[\exp\left(a_n\langle\gamma^u, \mathcal{P}Z_n\rangle\right)\right] + \tau - \langle\gamma^u, u\rangle \\
&\leq \frac{1}{a_n}\log\mathbb{E}\left[\exp\left(\langle a_n\mathcal{P}\gamma^u, Z_n\rangle\right)\right] + \tau - \langle\gamma^u, u\rangle \qquad (39) \\
&= \frac{1}{a_n}\Psi_n(a_n\mathcal{P}\gamma^u) + \tau - \langle\gamma^u, u\rangle,
\end{aligned}$$

where (39) 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

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

Recalling the properties of γ_i, we arrive at

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

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

Lemma 6.2. *Suppose that $D_\Psi\neq\{0\}$ is a subspace of $\mathbb{R}^p$, and $E\subseteq\mathbb{R}^p$ is closed and measurable. Then,*

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

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

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

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

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

$$
\begin{aligned}
&\leq \lim_{\zeta\to\infty} \limsup_{n\to\infty} \frac{1}{a_n} \log \sum_{j=1}^{\ell} \mu_n^j \left((-\infty,-\zeta]\right) + \mu_n^j \left([\zeta,\infty)\right) \\
&\leq \lim_{\zeta\to\infty} \limsup_{n\to\infty} \frac{1}{a_n} \log \left(2\ell \max_j \left\{ \mu_n^j \left((-\infty,-\zeta]\right), \mu_n^j \left([\zeta,\infty)\right) \right\} \right) \\
&= -\infty, \qquad (40)
\end{aligned}
$$

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 6.1 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 (40) yields the desired result. □

Theorem 6.1. *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 (11) as

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

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

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

because $\mathcal{P}$ is self-adjoint. Therefore, by Lemma 6.2,

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

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.

7 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 uniform tail probabilities for the pointwise estimation error. 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.

Acknowledgments

The second author's research was partially funded by the National Science Foundation (grant CMMI-2112828).

References

Adler, R. J. (2000), 'On excursion sets, tube formulas and maxima of random fields', *Annals of Applied Probability* **10**(1), 1–74.

Ankenman, B., Nelson, B. L. & Staum, J. (2010), 'Stochastic kriging for simulation metamodeling', *Operations Research* **58**(2), 371–382.

Arcones, M. A. (2006), 'Large deviations for M-estimators', *Annals of the Institute of Statistical Mathematics* **58**(1), 21–52.

Bect, J., Bachoc, F. & Ginsbourger, D. (2019), 'A supermartingale approach to Gaussian process based sequential design of experiments', *Bernoulli* **25**(4A), 2883–2919.

Beknazaryan, A., Sang, H. & Xiao, Y. (2019), 'Cramér type moderate deviations for random fields', *Journal of Applied Probability* **56**(1), 223–245.

Bull, A. D. (2011), 'Convergence rates of efficient global optimization algorithms', *Journal of Machine Learning Research* **12**, 2879–2904.

Chan, H. P. & Lai, T. L. (2006), 'Maxima of asymptotically Gaussian random fields and moderate deviation approximations to boundary crossing probabilities of sums of random variables with multidimensional indices', *The Annals of Probability* **34**(1), 80–121.

Cheng, D. & Xiao, Y. (2016), 'Excursion probability of Gaussian random fields on sphere', *Bernoulli* **22**(2), 1113–1130.

Ciesielski, Z. (1961), 'Hölder conditions for realizations of Gaussian processes', *Transactions of the American Mathematical Society* **99**(3), 403–413.

Dembo, A. & Zeitouni, O. (2009), *Large Deviations Techniques and Applications (2nd ed.)*, Springer Berlin Heidelberg.

Ghosal, S. & Roy, A. (2006), 'Posterior consistency of Gaussian process prior for nonparametric binary regression', *The Annals of Statistics* **34**(5), 2413–2429.

Glynn, P. W. & Juneja, S. (2004), A large deviations perspective on ordinal optimization, *in* R. Ingalls, M. D. Rossetti, J. S. Smith & B. A. Peters, eds, 'Proceedings of the 2004 Winter Simulation Conference', pp. 577–585.

Janson, S. (1987), 'Maximal spacings in several dimensions', *The Annals of Probability* **15**(1), 274–280.

Johnson, M. E., Moore, L. M. & Ylvisaker, D. (1990), 'Minimax and maximin distance designs', *Journal of Statistical Planning and Inference* **26**(2), 131–148.

Jones, D. R., Schonlau, M. & Welch, W. J. (1998), 'Efficient global optimization of expensive black-box functions', *Journal of Global Optimization* **13**(4), 455–492.

Joseph, V. R., Gul, E. & Ba, S. (2015), 'Maximum projection designs for computer experiments', *Biometrika* **102**(2), 371–380.

Lederer, A., Umlauft, J. & Hirche, S. (2019), Uniform error bounds for Gaussian process regression with application to safe control, *in* H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox & R. Garnett, eds, 'Advances in Neural Information Processing Systems', Vol. 32, Curran Associates, Inc., pp. 659–669.

Lee, S. I., Mortazavi, B., Hoffman, H. A., Lu, D. S., Li, C., Paak, B. H., Garst, J. H., Razaghy, M., Espinal, M., Park, E., Lu, D. C. & Sarrafzadeh, M. (2014), '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.

Li, J. & Ryzhov, I. O. (2022), 'Convergence rates of epsilon-greedy global optimization under radial basis function interpolation', *Stochastic Systems (to appear)*.

Li, X., Liu, J., Lu, J. & Zhou, X. (2018), 'Moderate deviation for random elliptic PDE with small noise', *The Annals of Applied Probability* **28**(5), 2781–2813.

Lukić, M. & Beder, J. (2001), 'Stochastic processes with sample paths in reproducing kernel Hilbert spaces', *Transactions of the American Mathematical Society* **353**(10), 3945–3969.

Marcus, M. B. & Shepp, L. A. (1972), Sample behavior of Gaussian processes, *in* L. M. Le Cam, J. Neyman & E. L. Scott, eds, 'Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability', Vol. 2, pp. 423–421.

Pati, D., Bhattacharya, A. & Cheng, G. (2015), 'Optimal Bayesian estimation in random covariate design with a rescaled Gaussian process prior', *Journal of Machine Learning Research* **16**, 2837–2851.

Pronzato, L. & Müller, W. G. (2012), 'Design of computer experiments: space filling and beyond', *Statistics and Computing* **22**(3), 681–701.

Rasmussen, C. E. & Williams, C. K. I. (2006), *Gaussian processes for machine learning*, MIT Press.

Scott, W. R., Powell, W. B. & Simão, H. P. (2010), Calibrating simulation models using the knowledge gradient with continuous parameters, *in* B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan & E. Yücesan, eds, 'Proceedings of the 2010 Winter Simulation Conference', pp. 1099–1109.

Sheibani, M. & Ou, G. (2021), 'The development of Gaussian process regression for effective regional post-earthquake building damage inference', *Computer-Aided Civil and Infrastructure Engineering* **36**(3), 264–288.

Snoek, J., Larochelle, H. & Adams, R. P. (2012), Practical Bayesian optimization of machine learning algorithms, *in* F. Pereira, C. J. C. Burges, L. Bottou & K. Q. Weinberger, eds, 'Advances in Neural Information Processing Systems', Vol. 25, pp. 2951–2959.

Srinivas, N., Krause, A., Kakade, S. M. & Seeger, M. W. (2012), 'Information-theoretic regret bounds for Gaussian process optimization in the bandit setting', *IEEE Transactions on Information Theory* **58**(5), 3250–3265.

Teckentrup, A. L. (2020), 'Convergence of Gaussian process regression with estimated hyper-parameters and applications in Bayesian inverse problems', *SIAM/ASA Journal on Uncertainty Quantification* **8**(4), 1310–1337.

Toth, C. & Oberhauser, H. (2020), Bayesian learning from sequential data using Gaussian processes with signature covariances, *in* H. Daumé & A. Singh, eds, 'Proceedings of the 37th International Conference on Machine Learning', pp. 9548–9560.

Vakili, S., Picheny, V. & Durrande, N. (2020), 'Regret bounds for noise-free Bayesian optimization', *arXiv preprint arXiv:2002.05096*.

van der Hofstad, R. & Honnappa, H. (2019), 'Large deviations of bivariate Gaussian extrema', *Queueing Systems* **93**(3), 333–349.

Vazquez, E. & Bect, J. (2010), 'Convergence properties of the expected improvement algorithm with fixed mean and covariance functions', *Journal of Statistical Planning and Inference* **140**(11), 3088–3095.

Wang, W., Tuo, R. & Wu, C. F. J. (2020), 'On prediction properties of kriging: Uniform error bounds and robustness', *Journal of the American Statistical Association* **115**(530), 920–930.

Wendland, H. (2004), *Scattered data approximation*, Cambridge University Press.

Wu, Z.-M. & Schaback, R. (1993), 'Local error estimates for radial basis function interpolation of scattered data', *IMA Journal of Numerical Analysis* **13**(1), 13–27.

Zhou, J. & Ryzhov, I. O. (2022), 'A new rate-optimal sampling allocation for linear belief models', *Operations Research (to appear)*.