Fast Statistical Leverage Score Approximation in Kernel Ridge Regression: Supplementary Materials

Yifan Chen
UIUC
yifanc10@illinois.edu

Yun Yang
UIUC
yy84@illinois.edu

The outline of the appendix is stated as follows. First, some useful results along with some proofs of the preliminaries and the main results are provided in Section 6. In Section 7 we report the detailed settings of the experiments in the main paper. In Section 8, we further analyze the properties of our leverage score approximation, which mirror the behavior of the true statistical leverage scores. To numerically accelerate the computation in the multivariate case, we simplify the multiple integral for obtaining the leverage score approximation to a single integral in Section 9. Besides, we show the error caused by density estimation is negligible compared with the total error in approximating statistical leverage scores in Section 10. Finally, we provide some technical facts regarding multivariate integration in Section 11.

6 USEFUL FACTS

6.1 Fourier Transform

Use $L_p(\mathbb{R}^d) = \{f : \mathbb{R}^d \to \mathbb{R}, \int_{\mathbb{R}^d} |f(x)|^p \, \mathrm{d}x < \infty\}$ to denote the space of all L_p integrable functions over $\mathbb{R}^d$ for $p \geq 1$. For any function $f \in L_1(\mathbb{R}^d)$, we use $\mathscr{F}[f]$ and $\mathscr{F}^{-1}[f]$ to denote its Fourier transform and its inverse Fourier transform, which is given by

$$\mathscr{F}[f](s) = \int_{\mathbb{R}^d} f(x)\, e^{-2\pi\sqrt{-1}\langle x,s\rangle} \, \mathrm{d}x, \quad \text{for all } s \in \mathbb{R}^d, \mathscr{F}^{-1}[f](x) = \int_{\mathbb{R}^d} f(s)\, e^{2\pi\sqrt{-1}\langle x,s\rangle} \, \mathrm{d}s, \quad \text{for all } x \in \mathbb{R}^d.$$

A useful property of the Fourier transform is the Parseval's identity.

Theorem 5 (Parseval's identity)**.** *For any $f \in L_2(\mathbb{R}^d)$, the following identity holds*

$$\int_{\mathbb{R}^d} |f(x)|^2 \, \mathrm{d}x = \int_{\mathbb{R}^d} |\mathscr{F}[f](s)|^2 \, \mathrm{d}s.$$

Besides, Fourier transform is closely related to kernels. For any PSD stationary kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, by Brochner's Theorem, it would be a Fourier transform of a Borel measure. For simplicity we abuse the notation by using $K(x)$ to mean $K(y, y + x), \forall y \in \mathbb{R}^d$, since $K(y, y + x)$ does not depend on the specific choice of y. Specifically, the Matérn kernel K_α with smoothness parameter $\nu = \alpha - d/2 > 0$ can be equivalently defined through its Fourier transform (Rasmussen, 2003, p. 84) as

$$m_\alpha(s) := \mathscr{F}[K_\alpha](s) = \int_{\mathbb{R}^d} K_\alpha(x)\, e^{-2\pi\sqrt{-1}\langle x,s\rangle} \, \mathrm{d}x = C_\alpha(1 + D_\alpha\|s\|^2)^{-\alpha}, \quad \forall s \in \mathbb{R}^d,$$

where C_α and D_α are some constants only dependent on α. (To simplify the statement of the theory, we would simply take $C_\alpha = D_\alpha = 1$ when later discussing the asymptotic properties.) Throughout this appendix, we focus on the case in which the Matérn kernel is used, since its theoretical properties have been well studied, and the proof is easy to be extended to other stationary kernels.

6.2 RKHS Associated with the Matérn Kernel—Proof of Theorem 1 in the Main Paper

The following theorem characterizes the RKHS $\mathbb{H}_\alpha$ associated with the Matérn kernel through its Fourier transform. The main body of the proof comes from the slides of Fukumizu (2008).

Lemma 6 (Fourier representation of RKHS). *For any $f, g \in \mathbb{H}_\alpha$, we have*

$$\|f\|^2_{\mathbb{H}_\alpha} = \int_{\mathbb{R}^d} \frac{\left|\mathscr{F}[f](s)\right|^2}{m_\alpha(s)}\,\mathrm{d}s, \quad \textit{and} \quad \langle f, g\rangle_{\mathbb{H}_\alpha} = \int_{\mathbb{R}^d} \frac{\mathscr{F}[f](s)\cdot\overline{\mathscr{F}[g](s)}}{m_\alpha(s)}\,\mathrm{d}s$$

Proof. Consider a measure space $(\mathbb{R}, \mathcal{B}, \mu)$, where $d\mu = m_\alpha(s)ds$, and $m_\alpha(s)$ is the spectral density of the invariant Matérn kernel with smoothness parameter $\nu = \alpha - \frac{d}{2}$. By referring to Section 3.1 in our main paper, we can check the function $m_\alpha(s) \in L^\infty$ is differentiable and positive everywhere. Based on the measure μ, we define a function space $\mathbb{G} = L^2(\mathbb{R}^d, \mu) \equiv \{F : \mathbb{R}^d \to \mathbb{C}; \int_{\mathbb{R}^d} |F|^2 d\mu < \infty\}$, with the inner product $\langle F, G\rangle_{\mathbb{G}} := \int F\overline{G}d\mu$. Here $\mathbb{C}$ is the set of all complex numbers. We can observe the form is quite similar to the frequency domain in Fourier Transform. To construct the RKHS of interest, we further define a function $H(s; x) = \exp(-2\pi\sqrt{-1}\,\langle x, s\rangle)$ and a map $\mathscr{M}(\cdot) : L^2(\mathbb{R}^d, \mu) \to \mathbb{T}$, similar to inverse Fourier Transform, given by

$$\mathscr{M}(F)(x) := \int F(s)\overline{H(s; x)}d\mu = \int F(s)\exp(2\pi\sqrt{-1}\,\langle x, s\rangle)\mathrm{d}\mu, \quad \forall x \in \mathbb{R}^d$$

where $\mathbb{T}$ is the space of all functions over $\mathbb{R}^d$ with the pointwise-convergence topology, i.e. $f_n \to f \Leftrightarrow f_n(x) \to f(x), \forall x \in \mathbb{R}^d$.

Now we are able to define a new function space $\mathbb{H} := \{f \in \mathbb{T};\, \exists F \in L^2(\mathbb{R}^d, \mu), f = \mathscr{M}(F)\}$, and equip it with the inner product $\langle f, g\rangle_{\mathbb{H}} := \langle F, G\rangle_{\mathbb{G}}$, where F, G satisfy $f = \mathscr{M}(F), g = \mathscr{M}(G)$. We first need to show $\mathbb{H}$ is an RKHS. We can check for all $f \in \mathbb{H}$,

$$f(x) = \langle F, H(\cdot; x)\rangle_{\mathbb{G}} = \langle f, \mathscr{M}(H(\cdot; x))\rangle_{\mathbb{H}}$$

Also, the reproducing kernel of $\mathbb{H}$ would be:

$$\begin{aligned} K(x, y) &= \langle \mathscr{M}(H(\cdot; x)), \mathscr{M}(H(\cdot; y))\rangle_{\mathbb{H}} = \langle H(\cdot; x), H(\cdot; y)\rangle_{\mathbb{G}} \\ &= \int \exp(-2\pi\sqrt{-1}\,\langle x, s\rangle)\exp(2\pi\sqrt{-1}\,\langle y, s\rangle)d\mu \\ &= \int \exp(2\pi\sqrt{-1}\,\langle y - x, s\rangle)m_\alpha(s)ds \\ &= K_\alpha(y - x) \end{aligned}$$

where K_α is the invariant Matérn kernel with smoothness parameter α. We can confirm $\mathbb{H}$ is exactly the RKHS induced by a Matérn kernel K_α.

To complete the proof, we still need to find the form of $\langle f, g\rangle_{\mathbb{H}}$. Note by the facts $m_\alpha(s) \in L^\infty$ and $F \in L^2$, we can infer $F(s)m_\alpha(s) \in L^2$ as $\int (F(s)m_\alpha(s))^2\mathrm{d}s \le \|m_\alpha\|^2_\infty \int F^2(s)\mathrm{d}s$. Using the Fourier isometry of L^2 (Adams and Fournier, 2003, Theorem 7.61), we obtain $F(s)m_\alpha(s) = \mathscr{F}[f](s)$; i.e., $F(s) = \mathscr{F}[f](s)/m_\alpha(s)$. Finally, by the definition of the inner product, for $f = \mathscr{M}(F)$ and $g = \mathscr{M}(G)$,

$$\begin{aligned} \langle f, g\rangle_{\mathbb{H}} = \langle F, G\rangle_{\mathbb{G}} &= \int \frac{\mathscr{F}[f](s)}{m_\alpha(s)}\frac{\overline{\mathscr{F}[g](s)}}{m_\alpha(s)}m_\alpha(s)ds \\ &= \int \frac{\mathscr{F}[f](s)\overline{\mathscr{F}[g](s)}}{m_\alpha(s)}ds \end{aligned}$$

in which the third equality relies on the fact the $m_\alpha(s)$ is real, and its conjugate is the same as itself. ♢

6.3 Embedding Inequalities

Let α be an integer. Following the notation of the book (Adams and Fournier, 2003), for any $p \ge 1$ and subset $\Omega \subset \mathbb{R}^d$, we use the notation $W^{\alpha,p}(\Omega)$ to denote the Sobolev space as a set of functions u in $L_p(\Omega)$ such that

u and its weak derivatives up to total order α have a finite L_p norm. With this definition, the Sobolev space admits a norm and a seminorm

$$\|u\|_{\alpha,p,\Omega} = \Big(\sum_{|\mathbf{k}|\le\alpha} \|D^{\mathbf{k}} u\|_{p,\Omega}^p \Big)^{\frac{1}{p}} = \Big(\sum_{|\mathbf{k}|\le\alpha} \int_\Omega \big|D^{\mathbf{k}} u(t)\big|^p \,\mathrm{d}t \Big)^{\frac{1}{p}}, \tag{1}$$

$$|u|_{\alpha,p,\Omega} = \Big(\sum_{|\mathbf{k}|=\alpha} \int_\Omega \big|D^{\mathbf{k}} u(t)\big|^p \,\mathrm{d}t \Big)^{\frac{1}{p}}. \tag{2}$$

where $\mathbf{k} = (k_1, \cdots, k_d)$ is a multi-index, and $D^{\mathbf{k}} u = \frac{\partial^{|\mathbf{k}|} u}{\partial^{k_1} x_1 \cdots \partial^{k_d} x_d}$. To more precisely describe the mixed derivative, we additionally define some notations here for future use:

- $\mathbf{A}$ is a subset of $[d]$, $-\mathbf{A} := ([d] - \mathbf{A})$,
- $u = \mathbf{I}_{\mathbf{A}} \in \{0,1\}^d$ is a vector s.t. $u_i = 1, \forall i \in \mathbf{A}; u_i = 0, \forall i \in -\mathbf{A}$,
- a vector $u = \mu(\mathbf{A}_{-1}, \mathbf{A}_{+1}) \in \{-1, 0, 1\}^d$ satisfies $u_i = -1, \forall i \in \mathbf{A}_{-1}; u_i = 1, \forall i \in \mathbf{A}_{+1}$,
- $x_{\mathbf{A}} := (x_i)_{i\in\mathbf{A}}$, and $g(x_{\mathbf{A}}; x_{-\mathbf{A}})$ represents $g(u)$ where $u_{\mathbf{A}} = x_{\mathbf{A}}$ are the variables and $u_{-\mathbf{A}} = x_{-\mathbf{A}}$ are taken as the parameters fixed in the integration,
- $C(y, \delta) := \{(x_1, x_2, \ldots, x_d); x_i \in [y_i - \delta, y_i + \delta], \forall i \in [d]\}$ is a cube centered at y,
- $C^{\mathbf{A}}(y, \delta) = C(y_{\mathbf{A}}, \delta)$ is the marginal cube of $C(y, \delta)$ defined as $\{x_{\mathbf{A}}; x_i \in (y_i - \delta, y_i + \delta), \forall i \in \mathbf{A}\}$.

We will primarily work with the case $p = 2$ and Ω being a connected domain. For any $u \in \mathbb{H}$, by using the fact that $\mathscr{F}[D^{\mathbf{k}} u](s) = (\prod_{i=1}^d (2\pi\sqrt{-1} s_i)^{k_i}) \cdot \mathscr{F}[u](s)$ and the Parseval's identity in Theorem 5, we obtain

$$\|u\|_{\alpha,2,\mathbb{R}^d}^2 = \int_{\mathbb{R}^d} \sum_{|\mathbf{k}|\le\alpha} \Big| \prod_{i=1}^d (2\pi\sqrt{-1} s_i)^{k_i} \, \mathscr{F}[u](s) \Big|^2 \,\mathrm{d}s.$$

Since there exist constants (C_1, C_2) such that $C_1(1 + \|s\|^2)^\alpha \le \big(\sum_{|\mathbf{k}|\le\alpha} \prod_{i=1}^d s_i^{k_i} \big)^2 \le C_2(1 + \|s\|^2)^\alpha$ holds for all $s \in \mathbb{R}^d$, by Lemma 6 we can further deduce that

$$C_1 \|u\|_{\mathbb{H}_\alpha}^2 \le \|u\|_{\alpha,2,\mathbb{R}^d}^2 \le C_2 \|u\|_{\mathbb{H}_\alpha}^2 \tag{3}$$

holds for any $u \in \mathbb{H}_\alpha$, the RKHS associated with the Matérn kernel with smoothness index $\nu (= \alpha - d/2)$.

We first invoke the following special case of interpolation theorem of Sobolev space $W^{\alpha,p}(\Omega)$ (Adams and Fournier, 2003, Theorem 5.12).

Theorem 7 (Interpolation inequality). *For any integer $0 \le k \le \alpha$, there exist two constants (c_0, K) only depending on α, such that for any $u \in W^{\alpha,2}(\Omega)$ and any $\varepsilon \in (0, c_0)$,*

$$|u|_{k,2,\Omega} \le K\big(\varepsilon^{\alpha-k} |u|_{\alpha,2,\Omega} + \varepsilon^{-k} \|u\|_{2,\Omega}\big),$$

where for any function g, $\|g\|_{2,\Omega}^2 = \int_\Omega g^2(t)\,\mathrm{d}t$.

We will also use the following generalization of Gagliardo–Nirenberg interpolation inequalities to bound the sup-norm, which can be viewed as an extension of the above interpolation inequality to the sup-norm.

Theorem 8 (Sup-norm interpolation inequality). *There exist two universal constants (c_1, K_1), such that for any $u \in W^{\alpha,2}(\Omega)$ and any $\varepsilon \in (0, c_1)$,*

$$\|u\|_{\infty,(1-\varepsilon^2)\Omega} := \sup_{t\in(1-\varepsilon^2)\Omega} |u(t)| \le K_1\big(\varepsilon^{-d} \|u\|_{2,\Omega} + \varepsilon^{2\alpha-d} |u|_{\alpha,2,\Omega}\big).$$

Proof. From the Gagliardo–Nirenberg interpolation inequalities (Brezis and Mironescu, 2018), we have

$$\|g\|_{\infty,\Omega} \le C(\|g\|_{2,\Omega} + |g|_{\alpha,\frac{d}{\alpha},\Omega}), \quad \forall\, g \in W^{\alpha,2}(\Omega),$$

where C is some universal constant. In fact, the last term could be further bounded by $C_0|g|_{\alpha,2,\Omega}$ since $W^{\alpha,2} = H_\alpha, \alpha > \frac{d}{2}, \frac{d}{\alpha} < 2$, and the domain Ω is bounded.

Now for each fixed point $y_0 \in (1-\varepsilon^2)\Omega$, we can obtain by the preceding display with $g(t) = u(y_0 + \varepsilon^2 t)$ in the above that

$$\|u\|_{\infty,B(y_0,\varepsilon^2)} \le \varepsilon^{-d}\,\|u\|_{2,B(y_0,\varepsilon^2)} + C_0\,\varepsilon^{2\alpha-d}\,|u|_{\alpha,2,B(y_0,\varepsilon^2)} \le C\varepsilon^{-d}\,\|u\|_{2,\Omega} + C_0\,\varepsilon^{2\alpha-d}\,|u|_{\alpha,2,\Omega},$$

where we have used the fact that $B(y_0,\varepsilon^2) \subset \Omega$ for any $y_0 \in (1-\varepsilon^2)\Omega$. Finally, the claimed inequality follows by the above derivation and the fact that

$$\|u\|_{\infty,(1-\varepsilon^2)\Omega} = \sup_{y_0 \in (1-\varepsilon^2)\Omega} \|u\|_{\infty,B(y_0,\varepsilon^2)}$$

♢

Theorem 7 and 8 leads to the following lemma that we will repeatedly use in our proof. Here we specifically consider a fixed point x_0 such that the density function $p(x)$ is uniformly bounded from below by $\frac{1}{2}\,p(x_0) > 0$ for any x satisfying $\|x - x_0\| < \delta(x_0)$ and some constant $\delta(x_0) > 0$ that may depend on x_0. Before entering the theorem, we define an RKHS norm $\|\cdot\|_\lambda$ as

$$\|f\|_\lambda^2 = \int_{\mathbb{R}^d} f^2(x)\,p(x)\,\mathrm{d}x + h^{2\alpha}\|f\|_{\mathbb{H}_\alpha}^2,$$

and its localized truncation $\|f\|_{x_0,\lambda}$ as

$$\|f\|_{x_0,\lambda}^2 := \int_{C(x_0,\delta(x_0))} f^2(x)\,p(x)\,\mathrm{d}x + h^{2\alpha}\|f\|_{\mathbb{H}_\alpha}^2.$$

Theorem 9 (Local interpolation for RKHS). *Suppose $\delta(x_0) \ge Ch\log(1/h)$ for some constant $C > 0$. Let $C(x_0,\delta(x_0))$ denote a cube centered at x_0 with edge length $2\delta(x_0)$. If $C\log(1/h) > c_0^{-1}$ and $\sqrt{C\log(1/h)} > c_1^{-1}$, where (c_0,c_1) are the constants in Theorems 7 and 8, then there exists a constant K' such that*

$$|f|_{k,2,C(x_0,\delta(x_0))} \le K' h^{-k} \max\left\{1, \big(p(x_0)\big)^{-1/2}\right\} \|f\|_{x_0,\lambda},$$

for any $f \in \mathbb{H}_\alpha$ and $k = 0,1,\ldots,\alpha$. In addition, for each $k = 0,1,\ldots\alpha$, there exists some constant $K'' > 0$ and $\varepsilon < c_1$ such that

$$\|f\|_{k,\infty,(1-\varepsilon^2)C(x_0,\delta(x_0))} \le K'' h^{-k-d/2} \max\left\{1, \big(p(x_0)\big)^{-1/2}\right\} \|f\|_{x_0,\lambda},$$

Finally, for the case $|\mathbf{A}| = k, |\mathbf{A}'| = k'$, and $\mathbf{A}' \subseteq \mathbf{A}$, there is also a constant K''' satisfying:

$$\Big(\int_{C^{\mathbf{A}}(x_0,\delta(x_0))} \big|D^{\mathbf{1}_{\mathbf{A}'}} f(x_{\mathbf{A}}; x_{-\mathbf{A}})\big|^2 \mathrm{d}x_{\mathbf{A}}\Big)^{\frac{1}{2}} \le K''' h^{-k'-(d-k)/2} \max\left\{1, \big(p(x_0)\big)^{-1/2}\right\} \|f\|_{x_0,\lambda},$$

which generalizes the first inequality and provides a finer L^2 norm control.

Proof. When $k = \alpha$, the first inequality is obvious due to the the equivalence (3) between $\|\cdot\|_{\alpha,2,\mathbb{R}}$ and $\|\cdot\|_{\mathbb{H}_\alpha}$, and the fact that $\|f\|_{x_0,\lambda} \ge h^\alpha \|f\|_{\mathbb{H}_\alpha}$. Now let us consider $k \le \alpha - 1$. Let $a = \delta(x_0)$ and $u(t) = f(x_0 + at)$ for $t \in \Omega$. By applying the change of variable formula for integral and the chain rule for derivatives, we obtain (with $x = at$)

$$\begin{aligned} |u|_{k,2,\Omega}^2 &= \sum_{|\mathbf{j}|=k} \int_\Omega \big|D^{\mathbf{j}} u(t)\big|^2\,\mathrm{d}t = \sum_{|\mathbf{j}|=k} a^{2k-d} \int_{C(x_0,a)} \big|D^{\mathbf{j}} f(x)\big|^2\,\mathrm{d}x \\ &= a^{2k-d}|f|_{k,2,C(x_0,a)}^2, \quad \forall\, k = 0,1,\ldots,\alpha. \end{aligned}$$

Combining this with Theorem 7 and the definition (1) of Sobolev norm yields

$$\begin{aligned}|f|^2_{k,2,C(x_0,\delta(x_0))} &\le a^{-(2k-d)}\bigg(K\varepsilon^{-k}\big(\varepsilon^{\alpha}|u|_{\alpha,2,\Omega}+\|u\|_{2,\Omega}\big)\bigg)^2\\ &\le 2K^2\,(a\varepsilon)^{-2k}\bigg((a\varepsilon)^{2\alpha}|f|^2_{\alpha,2,C(x_0,\delta(x_0))}+\int_{C(x_0,\delta(x_0))}\big|f(x)\big|^2\,\mathrm{d}x\bigg)\\ &\le 2K^2\,(a\varepsilon)^{-2k}\bigg((a\varepsilon)^{2\alpha}|f|^2_{\alpha,2,\mathbb{R}^d}+\int_{C(x_0,\delta(x_0))}\big|f(x)\big|^2\,\mathrm{d}x\bigg).\end{aligned}$$

Using the condition that $p(x)\ge p(x_0)/2$ for each $x\in C(x_0,\delta(x_0))$, and the equivalence (3) between $\|\cdot\|_{\alpha,2,\mathbb{R}^d}$ and $\|\cdot\|_{\mathbb{H}_\alpha}$, we further obtain by choosing $\varepsilon=h/\delta(x_0)\le (C\log(1/h))^{-1}<c_0$ in the above that

$$\begin{aligned}|f|^2_{k,2,C(x_0,\delta(x_0))} &\le 2K^2h^{-2k}\Big(h^{2\alpha}\,\|f\|^2_{\mathbb{H}_\alpha}+2\big(p(x_0)\big)^{-1}\int_{C(x_0,\delta(x_0))}\big|f(x)\big|^2\,p(x)\,\mathrm{d}x\Big)\\ &\le K'^2h^{-2k}\max\big\{1,\big(p(x_0)\big)^{-1}\big\}\|f\|^2_{x_0,\lambda},\end{aligned}$$

which yields the first claimed inequality.

To prove the second inequality, we will apply Theorem 8. More specifically, we apply Theorem 8 with $u(t)=D^{\mathbf{j}}f(x_0+at)$, $|\mathbf{j}|=k$, $a=\delta(x_0)$, $\Omega=C(x_0,\delta(x_0))$ and set $\varepsilon=\sqrt{h/a}\le\sqrt{\frac{1}{C\log(1/h)}}<c_1$ to obtain (with a change of variable formula for integration)

$$\begin{aligned}\|D^{\mathbf{j}}f\|_{\infty,(1-\varepsilon^2)\Omega}=\|u\|_{\infty,(1-\varepsilon^2)C(0,1)} &\le K_1\big(\varepsilon^{-d}\,\|u\|_{2,C(0,1)}+\varepsilon^{2\alpha-2k-d}|u|_{\alpha-k,2,C(0,1)}\big)\\ &\le K_1\varepsilon^{-d}a^{-d/2}\,|f|_{k,2,\Omega}+K_1\varepsilon^{2\alpha-2k-d}a^{\alpha-k-d/2}\,|f|_{\alpha,2,\Omega}\\ &= K_1h^{-d/2}\,|f|_{k,2,\Omega}+K_1h^{\alpha-k-d/2}\,|f|_{\alpha,2,\Omega}.\end{aligned}$$

Now, we can obtain by combining the above with the first inequality of this theorem,

$$|f|_{k,\infty,(1-\varepsilon^2)\Omega}\lesssim h^{-k-d/2}\max\big\{1,\big(p(x_0)\big)^{-1/2}\big\}\,\|f\|_{x_0,\lambda}.$$

To prove the final inequality, we will take $f(x_{\mathbf{A}};x_{-\mathbf{A}})$ as a k-d function. Analogously using the previous result, we have:

$$\int_{C^{\mathbf{A}}}\big|D^{\mathbf{1}_{\mathbf{A}'}}f(x_{\mathbf{A}};x_{-\mathbf{A}})\big|^2\mathrm{d}x_{\mathbf{A}}\le|f|^2_{k',2,C^{\mathbf{A}}}\le K'^2h^{-2k'}\big(\int_{C^{\mathbf{A}}}f^2(x_{\mathbf{A}};x_{-\mathbf{A}})\mathrm{d}x_{\mathbf{A}}+h^{2\alpha}|f(x_{\mathbf{A}};x_{-\mathbf{A}})|^2_{\alpha,2,C^{\mathbf{A}}}\big)$$

Then we take $D^{\mathbf{g}}f(x_{\mathbf{A}};x_{-\mathbf{A}})$ as a $(d-k)$-d function over $C^{-\mathbf{A}}(x_0,\delta(x_0))$ (g is any multi-index whose nonzero elements are in A'), and utilize the intermediate result of the second inequality:

$$\|D^{\mathbf{g}}f(x_{\mathbf{A}};x_{-\mathbf{A}})\|_{\infty,(1-\varepsilon^2)C^{-\mathbf{A}}}\lesssim h^{-(d-k)/2}|f|_{|\mathbf{g}|,2,C^{-\mathbf{A}}}+h^{-(d-k)/2+(\alpha-|\mathbf{g}|)}|f|_{\alpha,2,C^{-\mathbf{A}}}.$$

In that case,

$$\begin{aligned}\int_{C^{\mathbf{A}}}f^2(x_{\mathbf{A}};x_{-\mathbf{A}})\mathrm{d}x_{\mathbf{A}} &\lesssim \int_{C^{\mathbf{A}}}h^{-(d-k)}\|f(x_{\mathbf{A}};x_{-\mathbf{A}})\|^2_{2,C^{-\mathbf{A}}}+h^{-(d-k)+2\alpha}|f|^2_{\alpha,2,C^{-\mathbf{A}}}\mathrm{d}x_{\mathbf{A}}\\ &\lesssim h^{-(d-k)}\|f\|^2_{2,\Omega}+h^{-(d-k)+2\alpha}|f|^2_{\alpha,2,\Omega}\\ &\lesssim h^{-(d-k)}\max\big\{1,\big(p(x_0)\big)^{-1}\big\}\,\|f\|^2_{x_0,\lambda},\end{aligned}$$

The result for $h^{2\alpha}|f(x_{\mathbf{A}};x_{-\mathbf{A}})|^2_{\alpha,2,C^{\mathbf{A}}}$ could be analogously obtained. Combining the pieces together, we have

$$\int_{C^{\mathbf{A}}}\big|D^{\mathbf{1}_{\mathbf{A}'}}f(x_{\mathbf{A}};x_{-\mathbf{A}})\big|^2\mathrm{d}x_{\mathbf{A}}\lesssim h^{-2k'-(d-k)}\max\big\{1,\big(p(x_0)\big)^{-1}\big\}\,\|f\|^2_{x_0,\lambda}$$

♢

6.4 Leverage Score Approximation—Proof of Theorem 3 in the Main Paper

Proof. Let F denote the limiting cumulative distribution function of F_n, and p the density function associated with F. Recall that the rescaled leverage approximation $\widetilde{K}_\lambda(x, x_0)$ is the minimizer of the following local population level functional

$$A_{x_0}(f) = \frac{p(x_0)}{2}\int_{\mathbb{R}^d} f^2(x)\,\mathrm{d}x + \frac{\lambda}{2}\|f\|^2_{\mathbb{H}_\alpha} - f(x_0),$$

such that the following identity holds for each function $u \in \mathbb{H}_\alpha$, which corresponds to setting the Gateaux derivative DA_{x_0} of A_{x_0} at $\widetilde{K}_{x_0}$ to be the zero operator,

$$DA_{x_0}(\widetilde{K}_{x_0})(u) = p(x_0)\int_{\mathbb{R}^d} \widetilde{K}_{x_0}(x)\,u(x)\,\mathrm{d}x + \lambda\,\langle \widetilde{K}_{x_0},\, u\rangle_{\mathbb{H}_\alpha} - u(x_0) = 0,$$
$$\text{or}\quad \widetilde{K}_{x_0}(x) := \widetilde{K}_\lambda(x,\, x_0) = \mathscr{F}^{-1}\left[\frac{1}{p(x_0) + h^{2\alpha}\,(1+\|s\|^2)^\alpha}\right](x - x_0), \quad \forall x \in \mathbb{R}^d.$$

The rescaled leverage function G_{x_0} is instead the minimizer of the empirical functional A_{n,x_0}, and thus the Gateaux derivative DA_{n,x_0} at point G_{x_0} should be 0 since G_{x_0} is the optimal function for the functional. Using that fact,

$$\begin{aligned} DA_{n,x_0}(\widetilde{K}_{x_0})(\widetilde{u}) &= \{DA_{n,x_0}(\widetilde{K}_{x_0}) - DA_{n,x_0}(G(\cdot,\, x_0))\}(\widetilde{u}) \\ &= D^2 A_{n,x_0}(G(\cdot,\, x_0))(\widetilde{K}_{x_0} - G(\cdot,\, x_0), \widetilde{u}) \end{aligned}$$

The last equality holds due to the definition of second order functional derivative. Note the key identity that $D^2A_{n,x_0}(G(\cdot,\, x_0))(\widetilde{u}, \widetilde{u}) = \|\widetilde{u}\|^2_{n,\lambda}$. By choosing $u = \widetilde{u} := \widetilde{K}_{x_0} - G(\cdot,\, x_0)$, we would further have ($u, \widetilde{u}$ would be used interchangeably from now on)

$$DA_{n,x_0}(\widetilde{K}_{x_0})(u) = \|\widetilde{u}\|^2_{n,\lambda} = \int_{\mathbb{R}^d} \widetilde{u}^2(x)\,\mathrm{d}F_n(x) + \lambda\,\|\widetilde{u}\|^2_{\mathbb{H}_\alpha},$$

and our task somewhat reduces to bounding the term above $DA_{n,x_0}(u) = \|\widetilde{K}_{x_0} - G(\cdot,\, x_0)\|^2_{n,\lambda}$. To do that, we can expand the expression $DA_{n,x_0}(\widetilde{K}_{x_0})(u)$:

$$\begin{aligned} DA_{n,x_0}(\widetilde{K}_{x_0})(u) &= \int_{\mathbb{R}^d} \widetilde{K}_{x_0}(x)\,\mathrm{d}F_n(x) + \lambda\,\langle \widetilde{K}_{x_0},\, u\rangle_{\mathbb{H}_\alpha} - u(x_0) \\ &= \underbrace{DA_{x_0}(\widetilde{K}_{x_0})(u)}_{=0} + \underbrace{\int_{\mathbb{R}^d} \widetilde{K}_{x_0}(x)\,u(x)\,\mathrm{d}\big(F_n(x) - F(x)\big)}_{=:I_1} + \underbrace{\int_{\mathbb{R}^d} \widetilde{K}_{x_0}(x)\,u(x)\,\big(p(x) - p(x_0)\big)\,\mathrm{d}x}_{=:I_2}, \end{aligned}$$

and bound the last two terms separately.

Using Lemma 13, as $u = \widetilde{u}$ vanishes at infinity, we have

$$I_1 = (-1)^d \int_{\mathbb{R}^d} (F_n(x) - F(x))\frac{\partial^d}{\partial x_1 \partial x_2 \cdots \partial x_d}\big(\widetilde{K}_{x_0}(x)\,u(x)\big)\mathrm{d}x$$

The term $|I_1|$ can be correspondingly bounded as

$$\begin{aligned} |I_1| &\le \tau(n)\int_{\mathbb{R}^d} \Big|\frac{\partial^d}{\partial x_1 \partial x_2 \cdots \partial x_d}\big(\widetilde{K}_{x_0}(x)\,u(x)\big)\Big|\mathrm{d}x \\ &\le \tau(n)\sum_{\mathbf{k}_1 \sqcup \mathbf{k}_2 = [d]} \int_{\mathbb{R}^d} \big|D^{\mathbf{k}_1}\widetilde{K}_{x_0}(x) D^{\mathbf{k}_2} u(x)\big|\mathrm{d}x. \end{aligned}$$

Using Lemma 10(2) about the exponential decay on $\widetilde{K}_{x_0}$ and its derivatives and the local embedding inequalities

in Theorem 9, we obtain

$$\begin{aligned}\int_{\mathbb{R}^d}\left|D^{\mathbf{k}_1}\widetilde{K}_{x_0}(x)D^{\mathbf{k}_2}u(x)\right|\mathrm{d}x &\le \int_{C_{x_0,\,\delta(x_0)}}\left|D^{\mathbf{k}_1}\widetilde{K}_{x_0}(x)\right|\left|D^{\mathbf{k}_2}u(x)\right|\mathrm{d}x\\ &\quad+|u|_{|\mathbf{k}_2|,2,\mathbb{R}^d}\Big(\int_{C^c_{x_0,\,\delta(x_0)}}\left|D^{\mathbf{k}_1}\widetilde{K}_{x_0}(x)\right|^2\mathrm{d}x\Big)^{\frac{1}{2}}\\ \overset{(i)}{\le}&\Big(\int_{\mathbb{R}^d}\left|h^{-|\mathbf{k}_1|}(h^d+h^{-d})\,e^{-C_2\left\|x-x_0\right\|/h}\right|^2\mathrm{d}x\Big)^{1/2}\cdot|u|_{|\mathbf{k}_2|,2,C_{x_0,\delta(x_0)}}\\ &\quad+|u|_{|\mathbf{k}_2|,2,\mathbb{R}^d}\Big(\int_{\|x-x_0\|\ge Ch\log(1/h)}\left|h^{-|\mathbf{k}_1|}(h^d+h^{-d})\,e^{-C_2\left\|x-x_0\right\|/h}\right|^2\mathrm{d}x\Big)^{\frac{1}{2}}\end{aligned}$$

where step (i) follows by the Cauchy-Schwarz inequality and the assumption that $\delta(x_0)\ge Ch\log(1/h)$. Further bound is given as

$$\begin{aligned}&\int_{\mathbb{R}^d}\left|D^{\mathbf{k}_1}\widetilde{K}_{x_0}(x)D^{\mathbf{k}_2}u(x)\right|\mathrm{d}x\lesssim h^{-d/2-|\mathbf{k}_1|}\,|u|_{|\mathbf{k}_2|,2,C_{x_0,\delta(x_0)}}+\log^{\frac{d-1}{2}}(1/h)h^{C_2C-d/2-|\mathbf{k}_1|}\,|u|_{|\mathbf{k}_2|,2,\mathbb{R}^d}\\ \overset{(ii)}{\lesssim}&h^{-d/2-|\mathbf{k}_1|-|\mathbf{k}_2|}\,\max\left\{1,\big(p(x_0)\big)^{-1/2}\right\}\|u\|_{x_0,\lambda}+\log^{\frac{d-1}{2}}(1/h)h^{C_2C-d/2-|\mathbf{k}_1|}\,|u|_{|\mathbf{k}_2|,2,\mathbb{R}^d},\end{aligned}$$

where step (ii) uses the first inequality in Theorem 9 with $k=|\mathbf{k}_2|$. The next bound is derived as,

$$\begin{aligned}&\int_{\mathbb{R}^d}\left|D^{\mathbf{k}_1}\widetilde{K}_{x_0}(x)D^{\mathbf{k}_2}u(x)\right|\mathrm{d}x\\ \lesssim&h^{-3d/2}\,\max\left\{1,\big(p(x_0)\big)^{-1/2}\right\}\|u\|_{x_0,\lambda}+\log^{\frac{d-1}{2}}(1/h)h^{C_2C-d/2-|\mathbf{k}_1|}\,|u|_{|\mathbf{k}_2|,2,\mathbb{R}^d}\\ \lesssim&h^{-3d/2}\,\max\left\{1,\big(p(x_0)\big)^{-1/2}\right\}\|u\|_{x_0,\lambda}+\log^{\frac{d-1}{2}}(1/h)h^{C_2C-d/2-|\mathbf{k}_1|-\alpha}\|u\|_{x_0,\lambda},\end{aligned}$$

in which the last step utilizes the fact that $\|u\|_{1,2,\mathbb{R}}\le\|u\|_{\mathbb{H}_\alpha}\le h^{-\alpha}\|u\|_{x_0,\lambda}$. For $C>\alpha/C_2$, we can finally obtain

$$|I_1|\lesssim\tau(n)h^{-3d/2}\,\max\left\{1,\big(p(x_0)\big)^{-1/2}\right\}\|u\|_{x_0,\lambda}.$$

Similarly, by using the Lipschitz property of the density function p as $|p(x)-p(x_0)|\le\min\left\{2C_p,L_{x_0}\|x-x_0\|\right\}$ (where $C_p=\sup_x|p(x)|$ and L_{x_0} is the local Lipschitz constant of p around x_0), the exponential decay on $\widetilde{K}_{x_0}$ and the local embedding inequalities in Theorem 9 with $k=0$, we obtain

$$\begin{aligned}|I_2|&\lesssim\int_{C_{x_0,\,\delta(x_0)}}\left|\widetilde{K}_{x_0}(x)\right|\cdot\|x-x_0\|\cdot|u(x)|\,\mathrm{d}x+\|u\|_{\infty,\mathbb{R}^d}\int_{C^c_{x_0,\,\delta(x_0)}}\left|\widetilde{K}_{x_0}(x)\right|\mathrm{d}x\\ &\lesssim\Big(\int_{\mathbb{R}^d}\left|(h^d+h^{-d})\,e^{-C_2\left\|x-x_0\right\|/h}\,\|x-x_0\|\right|^2\mathrm{d}x\Big)^{1/2}\cdot\|u\|_{2,C_{x_0,\delta(x_0)}}\\ &\qquad+\|u\|_{\infty,\mathbb{R}^d}\int_{\|x-x_0\|\ge Ch\log(1/h)}\left|(h^d+h^{-d})\,e^{-C_2\left\|x-x_0\right\|/h}\right|\mathrm{d}x\\ &\lesssim h^{-d/2+1}\,\|u\|_{2,C_{x_0,\delta(x_0)}}+h^{C_2C}\,\|u\|_{\infty,\mathbb{R}^d}\lesssim h^{-d/2+1}\,\max\left\{1,\big(p(x_0)\big)^{-1/2}\right\}\|u\|_{x_0,\lambda}+h^{C_2C}\,\|u\|_{\infty,\mathbb{R}^d}.\end{aligned}$$

Putting pieces together, we obtain

$$\left|DA_{n,x_0}(\widetilde{K}_{x_0})(u)\right|\lesssim h^{C_2C}\,\|u\|_\infty+\max\left\{1,\big(p(x_0)\big)^{-1/2}\right\}\left(\tau(n)\,h^{-3d/2}+h^{-d/2+1}\right)\|u\|_{x_0,\lambda}.$$

Now we return back to the right hand side of the identity $DA_{n,x_0}(\widetilde{K}_{x_0})(u)=\|\widetilde{u}\|^2_{n,\lambda}$. Since F_n is nondecreasing, we have the following bound,

$$\begin{aligned}\int_{\mathbb{R}^d}\widetilde{u}^2(x)\,\mathrm{d}F_n(x)&\ge\int_{C(x_0,\delta(x_0))}\widetilde{u}^2(x)\,\mathrm{d}F_n(x)\\ &=\int_{C(x_0,\delta(x_0))}\widetilde{u}^2(x)\,\mathrm{d}F(x)+\int_{C(x_0,\delta(x_0))}\widetilde{u}^2(x)\,\mathrm{d}\big(F_n(x)-F(x)\big).\end{aligned}$$

Therefore, by the definition of the localized norm $\|\cdot\|_{x_0,\lambda}$, we have

$$\|\widetilde{u}\|_{n,\lambda}^2 \geq \|\widetilde{u}\|_{x_0,\lambda}^2 + \underbrace{\int_{C(x_0,\delta(x_0))} \widetilde{u}^2(x)\,\mathrm{d}\big(F_n(x)-F(x)\big)}_{=:I_3}.$$

By applying the Lemma 13 again (note $\widetilde{u}$ and $\widetilde{u}^2$ are infinitely differentiable), the second term I_3 can be bounded as (some terms are hidden)

$$\begin{aligned}|I_3| \lesssim \big\|\widetilde{u}^2(x)\big(F_n(x)-F(x)\big)\big\|_{\infty,C_{x_0,\delta(x_0)}} + \ldots \\ + \big\|F_n(x)-F(x)\big\|_{\infty,C_{x_0,\delta(x_0)}} \int_{C_{x_0,\delta(x_0)}} \big|\frac{\partial^d}{\partial x_1\partial x_2\cdots\partial x_d}(\widetilde{u}^2(x))\big|\,\mathrm{d}x.\end{aligned}$$

Now by applying the first and the second inequality in Theorem 9, and the Cauchy-Schwarz inequality, the sum of the two terms above can be bounded up to a constant by

$$\max\big\{1,\big(p(x_0)\big)^{-1}\big\}\,\tau(n)\,h^{-d}\,\|\widetilde{u}\|_{x_0,\lambda}^2.$$

Putting all the pieces together, we can reach

$$\begin{aligned}&\big(1-c\tau(n)\,\max\big\{1,\big(p(x_0)\big)^{-1}\big\}\,h^{-d}\big)\,\|\widetilde{u}\|_{x_0,\lambda}^2 \\ &\quad\leq c'\,h^{C_2C}\,\|\widetilde{u}\|_\infty + c'\,\max\big\{1,\big(p(x_0)\big)^{-1/2}\big\}\,\big(\tau(n)h^{-3d/2}+h^{-d/2+1}\big)\,\|\widetilde{u}\|_{x_0,\lambda}.\end{aligned}$$

It is easy to verify directly that we always have the crude bound $\|\widetilde{u}\|_\infty \lesssim n$, so by choosing constant C sufficiently large $h^{C_2C}n$ is decreasing, we can obtain from the above that

$$\|\widetilde{u}\|_{x_0,\lambda} \lesssim \max\big\{1,\big(p(x_0)\big)^{-1/2}\big\}\,\big(\tau(n)h^{-3d/2}+h^{-d/2+1}\big).$$

In addition, an application of the second inequality in Theorem 9 implies

$$\sup_{x\in C_{x_0,(1-h)\delta(x_0)}} |\widetilde{u}(x)| \lesssim \max\big\{1,\big(p(x_0)\big)^{-1/2}\big\}\,\big(\tau(n)\,h^{-2d}+h^{-d+1}\big).$$

Finally, by taking $x=y=x_0$ in the integral form of $\widetilde{K}_{x_0}$ in equation (4), we have the lower bound $\widetilde{K}_{x_0}(x_0)\geq ch^{-d}\big(p(x_0)\big)^{-1+1/(2\alpha)}$ for some constant $c>0$ that only depends on α. Therefore, we have the relative error bound

$$\frac{\big|\widetilde{K}_\lambda(x_0,x_0)-G(x_0,x_0)\big|}{\big|G(x_0,x_0)\big|} \lesssim \max\big\{1,\big(p(x_0)\big)^{1/2-1/(2\alpha)}\big\}\sqrt{p(x_0)}\,\big(\tau(n)\,h^{-d}+h\big),$$

for any x_0 such that the density function satisfies $p(x)\geq p(x_0)/2$ for all x in an $h\log(1/h)$ neighborhood of x_0. In particular, for any $\alpha\geq 1$, the relative error of estimating the leverage score remains bounded even if the local density $p(x_0)$ tends to zero. ♢

7 MORE ON SIMULATIONS

In this section, we mainly provide the complete experiment settings and one additional figure to help illustrate our method. We first describe all the competing methods: original kernel ridge regression; Nyström methods with uniform sampling (hereinafter referred to as "vanilla"); Nyström with Recursive-RLS (RC) (Musco and Musco, 2017); Nyström with BLESS (Rudi et al., 2018); and Nyström with spectral analysis (SA, our proposed method).

7.1 Experiment Settings in Figure 1 in the Main Paper

In this experiment, we compare the runtime and runtime versus error trade-off among Vanilla, RC, BLESS, and our method SA in Figure 1, under the 3-d bimodal setting ($\gamma = 0.4$) using the Matérn kernel ($\nu = 1.5$). Specifically, the bimodal distribution has two components: with probability $\frac{n}{n+n^\gamma}$ generating a Unif$[0,1]^3$; and with probability $\frac{n^\gamma}{n+n^\gamma}$ generating a random variable with pdf $\prod_{j=1}^3(5-2x_j)$ for $x_j \in [2, 2.5]$, where n is the sample size.

The sample size n ranges from $2,000$ to $500,000$. In particular, the target function is set as $f^*(x) = g(\|x\|_2/d)$ with $g(x) = 1.6|(x-0.4)(x-0.6)| - x(x-1)(x-2) - 0.5$, and i.i.d. noises follow $\mathcal{N}(0, 0.25)$; regularization parameter λ is set as $0.075 \cdot n^{-2/3}$, and the bandwidth for Gaussian kernel density estimator is $0.15n^{-1/7}$. The KDE estimator allows a 0.15 relative error. The projection dimension for all the methods is set as $5 \cdot n^{1/3}$, while the sub-sampling size s for all the iteration-based Nyström methods listed is chosen as $1 \cdot n^{1/3}$ due to high time complexity. All the results reported in Figure 1 are averaged over 30 replicates.

7.2 Experiment Settings in Table 1 in the Main Paper

Each method above is run on the `RadiusQueriesCount` (Savva et al., 2018; Anagnostopoulos et al., 2018)(denoted by RQP), `HTRU2` (Lyon et al., 2016), and `CCPP` (Tüfekci, 2014; Kaya and Tüfekci, 2012) datasets downloaded from the UCI ML Repository (Dua and Graff, 2017). Those datasets contain 10000, 17898, and 9568 data points, with 3, 8, and 5 features respectively. The smoothness parameter of Matérn kernel is set as $\nu = 0.5$, and $\alpha := \nu + \frac{d}{2} = \frac{d}{2} + 0.5$. The regularization parameter λ is set as $0.15 \cdot n^{-\frac{2\alpha}{2\alpha+d}}$. To attain the optimal error rate, the projection dimension of all methods $\lfloor 2 \cdot n^{\frac{d}{2\alpha+d}} \rfloor$; while the sub-sample size for estimating the statistical leverage scores in RC and BLESS is set as $\lfloor 1 \cdot n^{\frac{d}{2\alpha+d}} \rfloor$. We still use kernel density estimator to gain density estimation, and the detailed setting of this estimator is almost the same as the last experiment, using Gaussian kernel and the bandwidth $0.5 \cdot n^{-\frac{1}{3}}$. All the results reported in Table 1 are averaged over 10 replicates.

7.3 Experiment Settings in Figure 2 in the Main Paper

We ran the experiments on the one-dimensional (for the ease of visualization) Unif$[0,1]$, Beta$(15,2)$, and a bimodal distribution, as before, with two components: with probability $\frac{n}{n+n^\gamma}$ generating a Unif$[0,0.5]$; and with probability $\frac{n^\gamma}{n+n^\gamma}$ generating a random variable with pdf $(3-2x)$ for $x \in [1, 1.5]$, where n is the sample size and $\gamma = 0.6$. In addition, the Matérn kernel with smoothness parameter $\nu = 1.5$ is used, and density estimation is performed by a tree-based kernel density estimator. The number of observations varies from $n = 200$ to $10,000$. The regularization parameter of the KRR is set as $\lambda = 0.45 \cdot n^{-0.8}$.

A Gaussian kernel is used for density estimation, and the bandwidth is set to $1 \cdot n^{-0.2}$ for Uniform$[0,1]$ and $0.3 \cdot n^{-1/3}$ for the rest two distributions. Also, we allow a 0.05 relative error tolerance for density estimation since highly accurate density estimation is not required for Nyström methods (cf. Section 10). While implementing our algorithm, we also apply an ad-hoc modification to avoid the potential instability with a small density value $p(x_i)$, as mentioned in Section 3.1 in the main paper. Particularly, in the case of Beta distribution, if the density of point x_i is smaller than a threshold $h = 0.3 \cdot n^{-0.8}$, a weighted average $\frac{0.5h + p(x_i)}{1.5}$ would be used for the subsequent leverage score approximation.

In Figure 2, we show our method provides reasonably good approximations to the rescaled leverage scores across all settings. In particular, Unif$[0,1]$ is the easiest case (red curves) due to its flat density, which meets Assumption 1 and 2 for almost all design points; while for points with low density, such as those in the smaller cluster of the bimodal distribution and close to the boundary of Beta$(15,2)$, the absolute error tends to be large due to the leading constant C_{x_0} in the error bound in Theorem 3. Moreover, the relative approximation error has a clear tendency of decreasing as the sample size increases, which is also consistent with our theory.

7.4 The Additional Experiment for Gaussian Kernels

To show that our proposed method can also be extended to more kernels other than Matérn kernels, in this subsection we compare the in-sample prediction error among the methods above in Figure 3, under a dimension-increasing setting ($d = 3, 10, 30$ respectively) using a Gaussian kernel with bandwidth $\sigma = 1.5n^{-\frac{1}{2d+3}}$. We still

use a bimodal distribution similar to the above one: ($\gamma = 0.4$) with probability $\frac{n}{n+n^\gamma}$ generating a Unif$[0,1]^d$; and with probability $\frac{n^\gamma}{n+n^\gamma}$ generating a random variable with pdf $\prod_{j=1}^{d}(7-2x_j)$ for $x_j \in [3, 3.5]$, where n is the sample size.

The sample size n ranges from 1000 to 100,000. In particular, the target function is set as $f^*(x) = g(\|x\|_2/d) + g(x_1)$ (x_1 is the first element of x) with $g(x) = 1.6|(x-0.4)(x-0.6)| - x(x-1)(x-2) - 0.5$, and i.i.d. noises follow $\mathcal{N}(0, 0.25)$, which is the same as before; regularization parameter λ is set as $0.075 \cdot n^{-\frac{d+3}{2d+3}}$, and the bandwidth for the used Gaussian kernel density estimator is tuned for different dimension since when d is large, the density estimation will greatly fluctuate with the size of bandwidth. The projection dimension for all the methods is set as $5 \cdot n^{\frac{d}{2d+3}}$, while the sub-sampling size s for all the iteration-based Nyström methods listed is chosen as $1 \cdot n^{\frac{d}{2d+3}}$ due to high time complexity. All the results reported in Figure 1 are averaged over 20 replicates. From

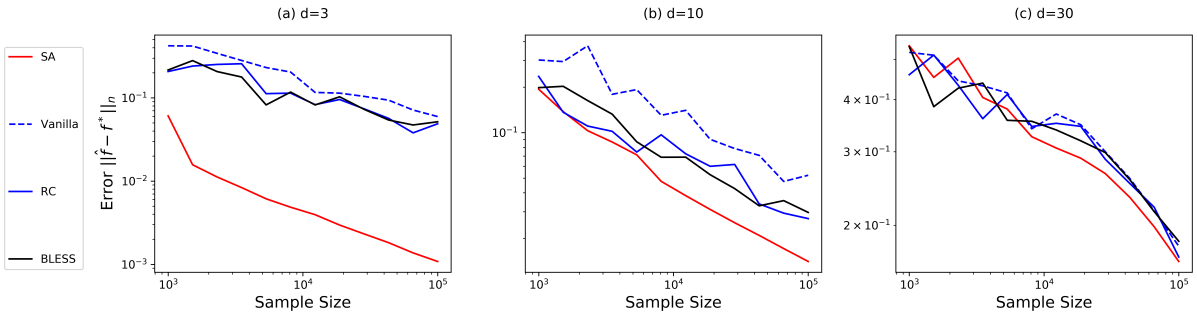

Figure 3: In-sample prediction error for Gaussian kernels with increasing dimension.

Figure 3, we observe when d increases, all the leverage-based methods will be no longer significantly better than vanilla uniform sampling, and the in-sample prediction error becomes orders of magnitude larger. We remark here that an increasing d indeed theoretically violates the assumption of kernel methods on the dimension. For the bad performance of KRR, we conjecture that is because in a high dimensional space the input samples get sparser (regarding the Euclidean distance), and thus roughly speaking for a certain sample with high density it is also hard to find some points around the sample, which is similar to the case for samples with low density.

8 APPROXIMATION PROPERTIES

In this section, we prove some useful properties of our equivalent kernel approximation introduced by Matérn kernels. Some parts of the proof rely on the isotropy of the stationary kernels. Since the isotropy is a property shared by most common stationary kernels, the proof is expected to be applied to other stationary kernels as well. For the reader's convenience, we also prove the corresponding lemmas for Gaussian kernels in Appendix 8.2. The proof strategy across the section is as follows, we first focus on one-dimensional cases and utilize the results to prove the conclusion for general multivariate approximation.

8.1 Matérn Kernel

For simplicity, we ignore some constants (such as $p(x_0)$ that does not change the local shape and scale of $\widetilde{K}_\lambda(\cdot, x_0)$) and instead consider the rescaled leverage approximation specified by

$$\widetilde{K}_\lambda(x, y) = \widetilde{K}_\lambda(x - y) = \int_{\mathbb{R}^d} \frac{e^{2\pi\sqrt{-1}\langle s, x-y\rangle}}{1 + \lambda\,(1 + \|s\|^2)^\alpha}\, ds = \int_{\mathbb{R}^d} \frac{\cos(2\pi\,\langle s, x-y\rangle)}{1 + \lambda\,(1 + \|s\|^2)^\alpha}\, ds, \tag{4}$$

where $\lambda = h^{2\alpha}$. By the inverse Fourier transform, we have

$$f_\lambda(s) = \frac{1}{1 + \lambda\,(1 + \|s\|^2)^\alpha} = \int_{\mathbb{R}^d} \widetilde{K}_\lambda(u)\, e^{-2\pi\sqrt{-1}\langle s, u\rangle}\, du. \tag{5}$$

Lemma 10. *When $2\alpha = 2\nu + d \geq d+1$ is an integer, we have:*

1. $\|\widetilde{K}_\lambda\|_\infty \lesssim h^{-d}$;
2. *There exists some constants $C_2 > 0$ such that*

$$|D^{\mathbf{j}} \widetilde{K}_\lambda(x,\, y)| \leq (h^{-|\mathbf{j}|-d})\, e^{-C_2\, \|x-y\|/h}, \quad |\mathbf{j}| = 0, 1, \ldots, d.$$

Proof. We start with the proof for the univariate case. From equation (4), we have

$$\|\widetilde{K}_\lambda\|_\infty \leq \int_{-\infty}^{\infty} \frac{1}{1+\lambda\,(1+s^2)^\alpha}\, ds \leq \int_{-\infty}^{\infty} \frac{1}{1+\lambda s^{2\alpha}}\, ds \lesssim \lambda^{-1/(2\alpha)} = h^{-1},$$

which is the first claimed property.

To prove the second property, we will apply the residue theorem to the following function

$$g(z) = \frac{e^{2\pi\sqrt{-1}|u|z}}{1+h^{2\alpha}\,(1+z^2)^\alpha}, \quad z \in \mathbb{C},$$

which is holomorphic on $\mathbb{C} \setminus \{z_1, \ldots, z_{2\alpha}\}$, where $z_1, \ldots, z_{2\alpha}$ are the 2α roots to the equation

$$1 + h^{2\alpha}\,(1+z^2)^\alpha = 0.$$

Therefore, z_{2k-1} and z_{2k}, for $k = 1, \ldots, \alpha$, are the two roots of the equation

$$z^2 = h^{-2}\, e^{\sqrt{-1}\frac{2k-1}{\alpha}\pi} - 1,$$

and $z_{2k-1} = -z_{2k}$. Without loss of generality, we assume $\mathrm{Im}(z_{2k-1}) > 0$. Direct calculations show that $|\,\mathrm{Im}(z_{2k-1})| \gtrsim h^{-1}$ and $|z_{2k-1}| \lesssim h^{-1}$ for each $k = 1, \ldots, \alpha$. Now we apply the residue theorem to the following contour integral

$$\int_C g(z)\, dz = \int_C \frac{e^{2\pi\sqrt{-1}|u|z}}{1+h^{2\alpha}\,(1+z^2)^\alpha}\, dz,$$

where the contour C goes along the real line from $-R$ to R and then counter-clockwise along a semicircle centering at 0 from R to $-R$, for some sufficiently large constant $R > 0$. The residue theorem implies

$$\int_C g(z)\, dz = 2\pi\sqrt{-1} \sum_{k=1}^{\alpha} \frac{e^{2\pi\sqrt{-1}|u|z_{2k-1}}}{2\alpha h^{2\alpha}(1+z_{2k-1}^2)^{\alpha-1} z_{2k-1}},$$

where we have used the fact that $\{z_{2k-1}\}_{k=1}^{\alpha}$ are the singularity points inside the contour C. Since $1 + h^{2\alpha}\,(1 + z_{2k-1}^2)^\alpha = 0$, the above can be further simplified into

$$\int_C g(z)\, dz = -\pi\sqrt{-1} \sum_{k=1}^{\alpha} \frac{e^{2\pi\sqrt{-1}|u|z_{2k-1}}(1+z_{2k-1}^2)}{\alpha z_{2k-1}}.$$

Due to the aforementioned properties that $|\,\mathrm{Im}(z_{2k-1})| \gtrsim h^{-1}$ and $|z_{2k-1}| \lesssim h^{-1}$, we have

$$\Big|\int_C g(z)\, dz\Big| \lesssim (h + h^{-1})\, e^{-C|u|/h}.$$

Finally, we can split the contour C into a straight part (real line) and a curved arc, so that

$$\int_C g(z)\, dz = \int_{(-R,R)} g(z)\, dz + \int_{\mathrm{arc}} g(z)\, dz,$$

where the arc part satisfies

$$\Big|\int_{\text{arc}} g(z)\,dz\Big| \le \pi R \cdot \sup_{\text{arc}} \Big|\frac{e^{2\pi\sqrt{-1}|u|z}}{1+h^{2\alpha}(1+z^2)^{\alpha}}\Big| \le \frac{\pi R}{h^{2\alpha}(R^2-1)^{\alpha}-1}.$$

By taking $R\to\infty$ (note that $\alpha>1/2$) and putting all pieces together, we finally reach

$$|\widetilde{K}_{\lambda}(x-y)| = \Big|\int_{-\infty}^{\infty} \frac{e^{2\pi\sqrt{-1}s(x-y)}}{1+\lambda\,(1+s^2)^{\alpha}}\,ds\Big| \lesssim (h+h^{-1})\,e^{-C|x-y|/h},$$

which is part of the second desired property.

To complete the proof of the second property, we still need to bound the derivative of the equivalent kernel. Recall the differentiation property of Fourier transform, and $\mathscr{F}[\widetilde{K}'_{\lambda}]$ could be written as:

$$\mathscr{F}[\widetilde{K}'_{\lambda}] = \frac{2\pi\sqrt{-1}s}{1+h^{2\alpha}\,(1+s^2)^{\alpha}}$$

Following a similar way, we can accordingly reset function g as:

$$g(z) = \frac{e^{2\pi\sqrt{-1}|u|z}2\pi\sqrt{-1}z}{1+h^{2\alpha}\,(1+z^2)^{\alpha}}, \quad z\in\mathbb{C},$$

and by the same procedure obtain the following equality

$$\int_C g(z)\,dz = 4\pi^2\sum_{k=1}^{\alpha}\frac{e^{2\pi\sqrt{-1}|u|z_{2k-1}}(1+z_{2k-1}^2)z_{2k-1}}{2\alpha z_{2k-1}} = 2\frac{\pi^2}{\alpha}\sum_{k=1}^{\alpha}e^{2\pi\sqrt{-1}|u|z_{2k-1}}(1+z_{2k-1}^2).$$

As for the integral over the arc part, its value is still negligible due to a finer control. Note over the arc, $z = R\cos(\theta)+\sqrt{-1}R\sin(\theta), \theta\in[0,\pi]$, and

$$\begin{aligned}\Big|\int_{\text{arc}} g(z)\,dz\Big| &= \Big|2\pi\sqrt{-1}\int_0^{\pi} e^{-2\pi|u|R\sin(\theta)}\frac{e^{2\pi\sqrt{-1}|u|R\cos(\theta)}z(\theta)}{1+h^{2\alpha}\,(1+z^2(\theta))^{\alpha}}(-R\sin(\theta)+\sqrt{-1}R\cos(\theta))\mathrm{d}\theta\Big| \\ &\le \frac{2\pi^2R^2}{h^{2\alpha}(R^2-1)^{\alpha}-1}\cdot 2\int_0^{\frac{\pi}{2}} e^{-2\pi|u|R\sin(\theta)}\mathrm{d}\theta.\end{aligned}$$

To bound the rest integral, we utilize the fact that $\sin(\theta)/\theta$ is decreasing in $(0,\pi/2]$, and $\sin(\theta)\ge\frac{2}{\pi}\theta, \forall\theta\in(0,\pi/2]$. Therefore,

$$\Big|\int_{\text{arc}} g(z)\,dz\Big| \le \frac{2\pi^2R^2}{h^{2\alpha}(R^2-1)^{\alpha}-1}\cdot 2\int_0^{\frac{\pi}{2}} e^{-4|u|R\theta}\mathrm{d}\theta \le \frac{C}{|u|}\frac{\pi^2R}{h^{2\alpha}(R^2-1)^{\alpha}-1}.$$

By taking $R\to\infty$ (note that $2\alpha>d=1$ here), the magnitude of the integral over the arc would vanish.

Putting all pieces together, we finally reach

$$|\widetilde{K}'_{\lambda}(x-y)| = \Big|\int_{-\infty}^{\infty}\frac{e^{2\pi\sqrt{-1}s(x-y)}(2\pi\sqrt{-1}s)}{1+\lambda\,(1+s^2)^{\alpha}}\,ds\Big| \lesssim (1+h^{-2})\,e^{-C|x-y|/h},$$

which complete the proof of the second property for univariate cases.

The proof for the multivariate claim will utilize the univariate conclusion before. By using the polar coordinate transform in Appendix 9, we can reduce the original multivariate integral to a univariate one (cf. equation (7)). The rescaled leverage $\widetilde{K}_{\lambda}(0)$ would be proportional to:

$$\int_0^{\infty}\frac{r^{d-1}}{1+\lambda(1+r^2)^{\alpha}}\mathrm{d}r.$$

which is of the scale h^{-d} by using the same technique as before. Therefore the first claim in this lemma has been proved.

For the second claim, we would heavily utilize the isotropy trick to simplify the proof. By the isotropy of Matérn kernels and our $\widetilde{K}_\lambda(u)$, we only need to consider a special input $\widetilde{u} = (\|u\|, 0, \ldots, 0)$. That is motivated by the observation that we can always do the coordinate transformation $s = T \cdot t$, where T is an orthogonal matrix and its first row $T_{1,\cdot} = u/\|u\|$. In that case, the original mixed derivative $D^{\mathbf{j}}\widetilde{K}_\lambda(u)$ could be expressed as

$$\int_{\mathbb{R}^d} \frac{e^{2\pi\sqrt{-1}\langle u,s\rangle}}{1+h^{2\alpha}\left(1+\|s\|^2\right)^\alpha} \prod_{i=1}^d (2\pi\sqrt{-1}s_i)^{\mathbf{j}_i} \mathrm{d}s = \int_{\mathbb{R}^d} \frac{e^{2\pi\sqrt{-1}\|u\|t_1}}{1+h^{2\alpha}\left(1+\|t\|^2\right)^\alpha} \prod_{i=1}^d (2\pi\sqrt{-1}\left\langle T_{i,\cdot}, t\right\rangle)^{\mathbf{j}_i} \mathrm{d}t,$$

which is of the same scale as $\max_{|\mathbf{j}'|=|\mathbf{j}|} |D^{\mathbf{j}'}\widetilde{K}_\lambda(\widetilde{u})|$. Under the settings above, the target considered would be reduced to

$$\begin{aligned}
&\Big|\int_{\mathbb{R}^d} \frac{e^{2\pi\sqrt{-1}\|u\|s_1}}{1+h^{2\alpha}\left(1+\sum_{i=1}^{d-1}s_i^2+s_d^2\right)^\alpha} \prod_{i=1}^d (s_i)^{\mathbf{j}_i} \mathrm{d}s\Big| = \Big|\int_{\mathbb{R}^{d-1}} \prod_{i=2}^d (s_i)^{\mathbf{j}_i} \int_{-\infty}^\infty \frac{e^{2\pi\sqrt{-1}\|u\|s_1} s_1^{\mathbf{j}_1}}{1+h^{2\alpha}\left(1+\|s_{-1}\|^2+s_1^2\right)^\alpha} \mathrm{d}s_1 \mathrm{d}s_{-1}\Big| \\
\le &\int_{\mathbb{R}^{d-1}} \prod_{i=2}^d |s_i|^{\mathbf{j}_i} \Big| \int_{-\infty}^\infty \frac{e^{2\pi\sqrt{-1}\|u\|s_1} s_1^{\mathbf{j}_1}}{1+h^{2\alpha}\left(1+\|s_{-1}\|^2+s_1^2\right)^\alpha} \mathrm{d}s_1\Big| \mathrm{d}s_{-1}.
\end{aligned}$$

The next important step is to take the expression $1+\|s_{-1}\|^2$ as a constant, and again apply the residue theorem to bound the internal integral as

$$\Big|\frac{\pi\sqrt{-1}}{\alpha}\sum_{k=1}^\alpha e^{2\pi\sqrt{-1}\|u\|z_{2k-1}}(1+z_{2k-1}^2)(z_{2k-1})^{\mathbf{j}_1-1}\Big| = \Big|\frac{\pi}{\alpha}\sum_{k=1}^\alpha e^{2\pi\sqrt{-1}\|u\|z_{2k-1}} h^{-2} e^{\sqrt{-1}\frac{2k-1}{\alpha}\pi}(z_{2k-1})^{\mathbf{j}_1-1}\Big|,$$

where $z_{2k-1} := a_k + \sqrt{-1}\cdot b_k$, and if we denote $\theta_k := \frac{2k-1}{\alpha}\pi$,

$$\begin{aligned}
a_k^2+b_k^2 = |z_{2k-1}^2| &= (h^{-4}\sin^2(\theta_k) + (h^{-2}\cos(\theta_k) - \|s_{-1}\|^2 - 1)^2)^{\frac{1}{2}}, \\
2b_k^2 = (a_k^2+b_k^2) - (a_k^2-b_k^2) &= (h^{-4}\sin^2(\theta_k) + (h^{-2}\cos(\theta_k) - \|s_{-1}\|^2 - 1)^2)^{\frac{1}{2}} - (h^{-2}\cos(\theta_k) - \|s_{-1}\|^2 - 1) \\
&= \frac{h^{-4}\sin^2(\theta_k)}{(h^{-4}\sin^2(\theta_k) + (h^{-2}\cos(\theta_k) - \|s_{-1}\|^2 - 1)^2)^{\frac{1}{2}} + (h^{-2}\cos(\theta_k) - \|s_{-1}\|^2 - 1)}
\end{aligned}$$

Note this time the magnitude of $\|s_{-1}\|$ matters a lot, and we need to divide the outside integral into two domains, $D_1 := \{|h^{-2}\cos(\theta_k) - \|s_{-1}\|^2 - 1| \le 3h^{-2}|\cos(\theta_k)|\}$ and $D_2 := \{|h^{-2}\cos(\theta_k) - \|s_{-1}\|^2 - 1| > 3h^{-2}|\cos(\theta_k)|\}$.

Over domain D_1, we can imply $\|s_{-1}\|^2 \le 4h^{-2}$, and similar to the univariate case we can have $b_k \gtrsim h^{-1}$ and $|z_{2k-1}| = \sqrt{a_k^2+b_k^2} = \Theta(h^{-1})$. The corresponding integral would be bounded by a constant multiple of

$$\begin{aligned}
&\int_{D_1} \prod_{i=2}^d |s_i|^{\mathbf{j}_i} h^{-2} e^{-C\|u\|b_k} (z_{2k-1})^{\mathbf{j}_1-1} \mathrm{d}s_{-1} \lesssim \int_{\|s_{-1}\|^2\le 4h^{-2}} \|s_{-1}\|^{|\mathbf{j}|-\mathbf{j}_1} h^{-2} e^{-C\|u\|h^{-1}} h^{-(\mathbf{j}_1-1)} \mathrm{d}s_{-1} \\
\lesssim &e^{-C\|u\|h^{-1}} h^{-(\mathbf{j}_1+1)} \int_0^{2h^{-1}} r^{|\mathbf{j}|-\mathbf{j}_1+d-2} \mathrm{d}r \lesssim e^{-C\|u\|h^{-1}} h^{-|\mathbf{j}|-d}.
\end{aligned}$$

For domain D_2, we could notice b_k would be much bigger than in D_1 as now $\|s_{-1}\|$ tends to dominate h^{-1}. Specifically, $2b_k^2 \ge 1+\|s_{-1}\|^2 + h^{-2}(1-\cos(\theta_k))$ and thus $b_k \ge C(\|s_{-1}\| + h^{-1})$. Considering $|z_{2k-1}| = \Theta(\|s_{-1}\|)$, we have

$$\begin{aligned}
&\int_{D_2} \prod_{i=2}^d |s_i|^{\mathbf{j}_i} h^{-2} e^{-C\|u\|b_k} (z_{2k-1})^{\mathbf{j}_1-1} \mathrm{d}s_{-1} \lesssim h^{-2} e^{-C\|u\|h^{-1}} \int_{\|s_{-1}\|^2 > c^2h^{-2}} \|s_{-1}\|^{|\mathbf{j}|-\mathbf{j}_1} e^{-C\|u\|\|s_{-1}\|} \|s_{-1}\|^{\mathbf{j}_1-1} \mathrm{d}s_{-1} \\
\lesssim &h^{-2} e^{-C\|u\|h^{-1}} \int_{ch^{-1}}^\infty r^{|\mathbf{j}|+d-3} e^{-C'\|u\|r} \mathrm{d}r \lesssim h^{-2} e^{-C\|u\|h^{-1}} h^{-|\mathbf{j}|-d+3} e^{-C'\|u\|h^{-1}} \lesssim e^{-C\|u\|h^{-1}} h^{-|\mathbf{j}|-d+1} \\
\lesssim &e^{-C\|u\|h^{-1}} h^{-|\mathbf{j}|-d}.
\end{aligned}$$

Combining the two pieces, we finally have

$$|D^{\mathbf{j}}\widetilde{K}_\lambda(u)| \lesssim e^{-C\|u\|h^{-1}} h^{-|\mathbf{j}|-d}.$$

♢

8.2 Gaussian Kernel

Similarly, we specify the equivalent kernel as

$$\widetilde{K}_\lambda(x,\, y) = \int_{\mathbb{R}^d} \frac{e^{2\pi\sqrt{-1}\langle s, x-y\rangle}}{1+\frac{\lambda}{\sigma^d}\, e^{\|\sigma s\|^2}}\, ds = \int_{\mathbb{R}^d} \frac{\cos(2\pi\,\langle s, x-y\rangle)}{1+\frac{\lambda}{\sigma^d}\, e^{\|\sigma s\|^2}}\, ds, \tag{6}$$

where the bandwidth σ is introduced due to its importance to exponential kernels. The scale of the bandwidth is set as $\mathcal{O}(\lambda^{\frac{1}{2\alpha}})$, where α is the corresponding parameter of the most suitable Matérn kernel attaining the optimal error rate in kernel ridge regression problems. We also define an auxiliary parameter h as $h^{-2} \equiv \ln\frac{\sigma^d}{\lambda}$ for simplicity of notation. In practice, σ would be specified in a way similar to Matérn case; a parameter $\alpha > d/2$ would be first chosen, and then let $\sigma = O(\lambda^{\frac{1}{2\alpha}}) \to 0$, which implies h here has the magnitude $O(\log^{-\frac{1}{2}}(n))$. In the following lemma, we will again start from a univariate case.

Lemma 11. *Given the equivalent kernel introduced above, we have ($\widetilde{\mathcal{O}}(\cdot)$ means $\mathcal{O}(\cdot)$ modulo poly-log terms):*

1. $\|\widetilde{K}_\lambda\|_\infty \lesssim \sigma^{-d}h^{-d} = \widetilde{\mathcal{O}}(\sigma^{-d})$;
2. *There exists some constant $C_3 > 0$ such that for $|\mathbf{j}| \le d$,*

$$|D^{\mathbf{j}}\widetilde{K}_\lambda(x,\, y)| \lesssim (\sigma h)^{-|\mathbf{j}|-d}\, e^{-C_3\,|x-y|\sigma^{-1}h} = \widetilde{\mathcal{O}}(\sigma^{-|\mathbf{j}|-d} e^{-C_3\,|x-y|\sigma^{-1}h}).$$

Proof. Again we begin with univariate cases. From equation (6), we have

$$\|\widetilde{K}_\lambda\|_\infty \le \int_{-\infty}^{\infty} \frac{1}{1+\frac{\lambda}{\sigma}\, e^{(\sigma s)^2}}\, ds \le 2\Big[\int_0^{\sigma^{-1}h^{-1}} \frac{1}{1+0}\, ds + \int_{\sigma^{-1}h^{-1}}^{\infty} \frac{1}{\frac{\lambda}{\sigma}\, e^{(\sigma s)^2}}\, ds\Big]$$

The integral is divided into two parts in which 1 and $\frac{\lambda}{\sigma}\, e^{(\sigma s)^2}$ dominate respectively. What's more, applying the property of error function that $\int_x^\infty e^{-t^2}\, dt = \mathcal{O}(\frac{e^{-x^2}}{x})$ (it holds when x is large enough), we can further obtain

$$\|\widetilde{K}_\lambda\|_\infty \lesssim 2\Big[\sigma^{-1}h^{-1} + \frac{1}{\lambda}he^{-h^{-2}}\Big] \lesssim \sigma^{-1}h^{-1}$$

which is the first claimed property.

To prove the second one, we still apply the residue theorem to the following function ($|x-y|$ is denoted as $|u|$ for simplicity),

$$g(z) = \frac{e^{2\pi\sqrt{-1}|u|z}}{1+\frac{\lambda}{\sigma}\, e^{(\sigma z)^2}}, \quad z \in \mathbb{C},$$

which is holomorphic on the complex plane except the roots $z_i, i \in \mathbb{Z}$ to the following equation:

$$1 + \frac{\lambda}{\sigma}\, e^{(\sigma z)^2} = 0$$

Therefore, z_{2k-1} and z_{2k}, for $k \in \mathbb{Z}$, are the two roots of the equation:

$$\sigma^2 z^2 = (h^{-2} + \sqrt{-1}(2k-1)\pi)$$

and $z_{2k-1} = -z_{2k}$. Without loss of generality, we assume $\mathrm{Im}(z_{2k-1}) > 0$. Direct calculations roughly show that $|\mathrm{Im}(z_{2k-1})| \gtrsim \sigma^{-1}$ and $|z_{2k-1}| \lesssim \sigma^{-1}$ for each $k \in \mathbb{Z}$. Further analysis would be provided later.

Note we could only focus on the case $|u| > \sigma h^{-1}$, since otherwise

$$|\widetilde{K}_\lambda(u)| \le C(\sigma h)^{-1} \le C\, e^{C_3}(\sigma h)^{-1} e^{-C_3|u|\sigma^{-1}h},$$

and the claimed property would be proved automatically. Now we apply the residue theorem to the following contour integral

$$\int_C g(z)\, dz = \int_C \frac{e^{2\pi\sqrt{-1}|u|z}}{1 + \frac{\lambda}{\sigma}\, e^{(\sigma z)^2}}\, dz,$$

where the contour C goes along the real line from $-R$ to R and then counter-clockwise along a semicircle centering at 0 from R to $-R$, for some sufficiently large constant $R > 0$. Denote the index set A_R as the set of all the indices k that the roots $\{z_{2k-1}\}_k$ are inside the contour C. The residue theorem implies

$$\int_C g(z)\, dz = 2\pi\sqrt{-1} \sum_{k \in A_R} \frac{e^{2\pi\sqrt{-1}|u|z_{2k-1}}}{2\sigma\lambda e^{(\sigma z_{2k-1})^2} z_{2k-1}},$$

Since $1 + \frac{\lambda}{\sigma}\, e^{(\sigma z)^2} = 0$, the above expression can be further simplified into

$$\int_C g(z)\, dz = -\pi\sqrt{-1} \sum_{k \in A_R} \frac{e^{2\pi\sqrt{-1}|u|z_{2k-1}}}{\sigma^2 z_{2k-1}}.$$

Note the set A_R is symmetric about 0 and goes to $\mathbb{Z}$ as $R \to \infty$. We can first pair the opposite k and denote $z_{2k-1} = a_k + b_k\sqrt{-1}, z_{1-2k} = -a_k + b_k\sqrt{-1}$ for convenience. In this case,

$$\begin{aligned}
&\frac{e^{2\pi\sqrt{-1}|u|z_{2k-1}}}{\sigma^2 z_{2k-1}} + \frac{e^{2\pi\sqrt{-1}|u|z_{1-2k}}}{\sigma^2 z_{1-2k}} = \frac{e^{-2\pi|u|b_k}}{-\sigma^2(a_k^2 + b_k^2)} 2\sqrt{-1}(b_k \cos(2\pi|u|a_k) - a_k \sin(2\pi|u|a_k)) \\
=&\frac{2\sqrt{-1}e^{-2\pi|u|b_k}}{-\sigma^2\sqrt{a_k^2 + b_k^2}} \cos(2\pi|u|a_k + \arctan(a_k/b_k))
\end{aligned}$$

And hence the sequence of the integral over the semicircle C with radius R would converge to

$$\frac{-2\pi}{\sigma^2} \sum_{k=1}^{\infty} \frac{e^{-2\pi|u|b_k}}{\sqrt{a_k^2 + b_k^2}} \cos(2\pi|u|a_k + \arctan(a_k/b_k)) \le \frac{2\pi}{\sigma^2} \sum_{k=1}^{\infty} \frac{e^{-2\pi|u|b_k}}{\sqrt{a_k^2 + b_k^2}}$$

To further analyze the scale of the infinite series, we need to uncover the form of the coefficient a_k, b_k. Recall $z_{2k-1}^2 = \frac{1}{\sigma^2}(\ln\frac{\sigma}{\lambda} + (2k-1)\pi)$, and the corresponding derivation is,

$$\begin{aligned}
a_k^2 + b_k^2 &= \sigma^{-2}(\ln^2\frac{\sigma}{\lambda} + ((2k-1)\pi)^2)^{\frac{1}{2}} \\
2b_k^2 &= (a_k^2 + b_k^2) - (a_k^2 - b_k^2) = \sigma^{-2}\Big[(\ln^2\frac{\sigma}{\lambda} + ((2k-1)\pi)^2)^{\frac{1}{2}} - (\ln\frac{\sigma}{\lambda})\Big] \\
&= \frac{\sigma^{-2}((2k-1)\pi)^2}{(\ln^2\frac{\sigma}{\lambda} + ((2k-1)\pi)^2)^{\frac{1}{2}} + (\ln\frac{\sigma}{\lambda})}
\end{aligned}$$

Denote $H \equiv (\frac{h^{-2}}{\pi} + 1)/2$. The curve of a_k, b_k could be roughly divided into two stages, $k \le \lfloor H \rfloor$ and $k \ge \lceil H \rceil$. In the first stage, $a_k^2 + b_k^2 \ge \sigma^{-2}\ln\frac{\sigma}{\lambda} = (\sigma h)^{-2}$ and $2b_k^2 \ge \sigma^{-2}\frac{((2k-1)\pi)^2}{2\ln\frac{\sigma}{\lambda}} = \sigma^{-2}\frac{((2k-1)\pi)^2}{2h^{-2}}$; in the second stage, $a_k^2 + b_k^2 \ge \sqrt{2}\pi\sigma^{-2}(2k-1)$ and $2b_k^2 \ge \sigma^{-2}\frac{((2k-1)\pi)^2}{3\pi(2k-1)} = \sigma^{-2}\frac{(2k-1)\pi}{3}$. The infinite series could be therefore bounded as

$$\begin{aligned}
\frac{2\pi}{\sigma^2} \sum_{k=1}^{\lfloor H \rfloor} \frac{e^{-2\pi|u|b_k}}{\sqrt{a_k^2 + b_k^2}} &\lesssim \sigma^{-1} h \sum_{k=1}^{\lfloor H \rfloor} e^{-C_3|u|\sigma^{-1}hk} \\
\frac{2\pi}{\sigma^2} \sum_{k=\lceil H \rceil}^{\infty} \frac{e^{-2\pi|u|b_k}}{\sqrt{a_k^2 + b_k^2}} &\lesssim \sigma^{-2} \sum_{k=\lceil H \rceil}^{\infty} \frac{e^{-C_3|u|\sigma^{-1}\sqrt{k}}}{\sigma^{-1}\sqrt{k}}
\end{aligned}$$

The two series above will converge rapidly when n is large enough. The scale of the two series above will therefore depend on their own first terms. The first series would be bounded as a constant multiple of $\sigma^{-1}h \cdot h^{-2}e^{-C_3|u|\sigma^{-1}h\cdot 1} = (\sigma h)^{-1}e^{-C_3|u|\sigma^{-1}h}$; due to the decreasing sequence, the last series could be bounded in the following way (note $|u| > \sigma h^{-1}$),

$$\begin{aligned}\sigma^{-1}\sum_{k=\lceil H\rceil}^{\infty}\frac{e^{-C_3|u|\sigma^{-1}\sqrt{k}}}{\sqrt{k}} &\le \sigma^{-1}\Big[\frac{e^{-C_3|u|\sigma^{-1}\sqrt{\lceil H\rceil}}}{\sqrt{\lceil H\rceil}} + \int_{\lceil H\rceil}^{\infty}\frac{e^{-C_3|u|\sigma^{-1}\sqrt{x}}}{\sqrt{x}}d\,x\Big]\\ &\le \sigma^{-1}\Big[he^{-C_3|u|(\sigma h)^{-1}} + \frac{2\sigma}{C_3|u|}e^{-C_3|u|(\sigma h)^{-1}}\Big] \lesssim \sigma^{-1}he^{-C_3|u|(\sigma h)^{-1}}\end{aligned}$$

And hence the whole series would be bounded by $(\sigma h)^{-1}e^{-C_3|u|\sigma^{-1}h}$.

Then, we split the contour C into a straight part (real line) and a curved arc, so that

$$\int_C g(z)\,dz = \int_{(-R,R)} g(z)\,dz + \int_{\text{arc}} g(z)\,dz,$$

where the arc part satisfies $z = Re^{\sqrt{-1}\theta}, \theta \in [0,\pi]$ and hence,

$$\Big|\int_{\text{arc}} g(z)\,dz\Big| = \Big|\int_0^{\pi}\frac{e^{2\pi\sqrt{-1}|u|Re^{\sqrt{-1}\theta}}}{1+\frac{\lambda}{\sigma}e^{\sigma^2R^2e^{2\sqrt{-1}\theta}}}\sqrt{-1}Re^{\sqrt{-1}\theta}\mathrm{d}\theta\Big|$$

where the module of the integrand could be bounded by $\frac{e^{-2\pi|u|R\sin\theta}R}{|1-\frac{\lambda}{\sigma}e^{\sigma^2R^2\cos(2\theta)}|}$. By taking $R\to\infty$ and requiring $|u|>0$, we could observe that when $\sin\theta$ is bounded away from 0 the integrand is exponentially decaying; when $\sin\theta$ is nearly zero $\cos 2\theta = 1-2\sin^2\theta \gg 0$ and the integrand would also go to 0. That's to say, the whole integral $\int_{\text{arc}} g(z)\,dz \to 0$. Putting all pieces together, we finally reach

$$|\widetilde{K}_\lambda(x-y)| = \int_{-\infty}^{\infty}\frac{e^{2\pi\sqrt{-1}s(x-y)}}{1+\frac{\lambda}{\sigma}e^{(\sigma s)^2}}\,ds \lesssim \sigma^{-1}he^{-C_3|u|\sigma^{-1}h} + \sigma^{-1}h^{-1}e^{-C_3|u|\sigma^{-1}h^{-1}},$$

which is part of the second desired property and similar to the conclusion in (10).

To complete the proof of the second property, we still need to bound the derivative of the equivalent kernel. Recall the differentiation property of Fourier transform, and $\mathscr{F}[\widetilde{K}'_\lambda]$ could be written as:

$$\mathscr{F}[\widetilde{K}'_\lambda] = \frac{2\pi\sqrt{-1}s}{1+\frac{\lambda}{\sigma}e^{(\sigma s)^2}}.$$

With the expression above we can bound the sup norm of the derivative,

$$\|\widetilde{K}'_\lambda\|_\infty \le 4\pi\int_0^{\infty}\frac{s}{1+\frac{\lambda}{\sigma}e^{(\sigma s)^2}}\,ds \le 4\pi\Big[\int_0^{\sigma^{-1}h^{-1}}\frac{s}{1+0}\,ds + \int_{\sigma^{-1}h^{-1}}^{\infty}\frac{s}{\frac{\lambda}{\sigma}e^{(\sigma s)^2}}\,ds\Big].$$

The integral is divided into two parts in which 1 and $\frac{\lambda}{\sigma}e^{(\frac{s}{\sigma})^2}$ dominate respectively. We can further obtain

$$\|\widetilde{K}'_\lambda\|_\infty \lesssim \Big[(\sigma h)^{-2} + \frac{1}{2\lambda\sigma}e^{-h^{-2}}\Big] \lesssim (\sigma h)^{-2}$$

To exactly analyze behavior of the derivative of equivalent kernels, we can accordingly reset function g as:

$$g(z) = \frac{e^{2\pi\sqrt{-1}|u|z}2\pi\sqrt{-1}z}{1+\frac{\lambda}{\sigma}e^{(\sigma z)^2}}, \quad z\in\mathbb{C},$$

and by the same procedure obtain the following inequality:

$$\begin{aligned}\int_C g(z)\,dz &= 2\pi^2\sqrt{-1}\sum_{k\in A_R}\frac{e^{2\pi\sqrt{-1}|u|z_{2k-1}}}{\sigma^2} = \frac{2\pi^2}{\sigma^2}\sum_{k=1}^{\infty}(e^{2\pi\sqrt{-1}|u|(a_k+b_k\sqrt{-1})} + e^{2\pi\sqrt{-1}|u|(-a_k+b_k\sqrt{-1})})\\ &= \frac{4\pi^2}{\sigma^2}\sum_{k=1}^{\infty}e^{-2\pi|u|b_k}(\cos(2\pi|u|a_k)) \le \frac{4\pi^2}{\sigma^2}\sum_{k=1}^{\infty}e^{-2\pi|u|b_k}.\end{aligned}$$

We would only focus on the case $|u| > (\sigma/h)^{-1}$, since otherwise $|\widetilde{K}'_\lambda(u)| \le C(\sigma h)^{-2}$, which is bounded by $\le C\, e^{C_3}(\sigma h)^{-2} e^{-C_3|u|\sigma^{-1}h}$, and the claimed property would be proved automatically. Due to the aforementioned division of the series, we have (note $|u| > (\sigma/h)^{-1}$),

$$\begin{aligned}
\int_C g(z)\,dz &\le \frac{4\pi^2}{\sigma^2}\Big(\sum_{k=1}^{\lfloor H\rfloor} e^{-2\pi|u|b_k} + \sum_{k=\lceil H\rceil}^{\infty} e^{-2\pi|u|b_k}\Big)\\
&\lesssim \frac{4\pi^2}{\sigma^2}\Big[h^{-2}e^{-C_3|u|\sigma^{-1}h} + e^{-C_3|u|\sigma^{-1}h^{-1}} + \int_{k=\lceil H\rceil}^{\infty} e^{-C_3|u|\sigma^{-1}\sqrt{k}}d\,k\Big]\\
&\lesssim \frac{1}{(\sigma h)^2}\Big[e^{-C_3|u|\sigma^{-1}h} + \frac{2\sigma}{C_3|u|}\Big(h^{-2}e^{-C_3|u|\sigma^{-1}h^{-1}} + \frac{\sigma}{C_3|u|}e^{-C_3|u|\sigma^{-1}h^{-1}}\Big)\Big]\\
&\lesssim \frac{1}{(\sigma h)^2}e^{-C_3|u|\sigma^{-1}h}.
\end{aligned}$$

As for the integral over the arc part, its value is still negligible as

$$\Big|\int_{\text{arc}} g(z)\,dz\Big| \le \Big|\int_0^{\pi} \frac{e^{2\pi\sqrt{-1}|u|Re^{\sqrt{-1}\theta}}2\pi\sqrt{-1}Re^{\sqrt{-1}\theta}}{1+\frac{\lambda}{\sigma}e^{\sigma^2R^2e^{2\sqrt{-1}\theta}}}R\,d\theta\Big|$$

where the module of the integrand could be bounded by $\frac{e^{-2\pi|u|R\sin\theta}2\pi R^2}{|1-\frac{\lambda}{\sigma}e^{\sigma^2R^2\cos(2\theta)}|}$. The bound goes to 0 when $R\to\infty$ and $|u|>0$ are assumed as before, and the integral over the arc is again negligible. Putting all pieces together, we finally reach

$$|\widetilde{K}'_\lambda(x-y)| = \Big|\int_{-\infty}^{\infty}\frac{e^{2\pi\sqrt{-1}s(x-y)}(2\pi\sqrt{-1}s)}{1+\frac{\lambda}{\sigma}e^{(\sigma s)^2}}\,ds\Big| \lesssim (\sigma h)^{-2}\,e^{-C_3|u|\sigma^{-1}h},$$

which completes the proof of the second property in univariate cases.

For multivariate cases, again by applying polar coordinate transformation, the original integral could be bounded as

$$\begin{aligned}
\|\widetilde{K}_\lambda\|_\infty &\lesssim \int_0^{\infty}\frac{r^{d-1}}{1+\frac{\lambda}{\sigma^d}e^{(\sigma r)^2}}\,\mathrm{d}r \le \int_0^{\sigma^{-1}h^{-1}}\frac{r^{d-1}}{1+0}\mathrm{d}r + \int_{\sigma^{-1}h^{-1}}^{\infty}\frac{r^{d-1}}{\frac{\lambda}{\sigma^d}e^{(\sigma r)^2}}\,dr\\
&\lesssim (\sigma h)^{-d} + \frac{1}{\lambda}\int_{h^{-1}}^{\infty} r^{d-1}e^{-r^2}\mathrm{d}r \lesssim (\sigma h)^{-d} - \frac{1}{\lambda}\int_{h^{-1}}^{\infty} r^{d-2}\mathrm{d}e^{-r^2}.
\end{aligned}$$

After repeatedly using integration by parts, the last term could be further bounded by $(\sigma h)^{-d} + \sigma^{-d}h^{-(d-2)}$, which validates the first claim in this lemma.

For the second claim, we would still use the same strategy, utilizing the isotropy and some other tricks, as in the proof for Matérn kernels. Specifically, we would focus on the special case

$$\Big|\int_{\mathbb{R}^d}\frac{e^{2\pi\sqrt{-1}\|u\|s_1}}{1+\frac{\lambda}{\sigma^d}e^{\sigma^2(\|s_{-1}\|^2+s_1^2)}}\prod_{i=1}^{d}s_i^{\mathbf{j}_i}\mathrm{d}s\Big| \le \int_{\mathbb{R}^{d-1}}\prod_{i=2}^{d}|s_i|^{\mathbf{j}_i}\cdot\Big|\int_{-\infty}^{\infty}\frac{e^{2\pi\sqrt{-1}\|u\|s_1}|s_1|^{\mathbf{j}_1}}{1+\frac{\lambda}{\sigma^d}e^{\sigma^2(\|s_{-1}\|^2+s_1^2)}}\mathrm{d}s_1\Big|\mathrm{d}s_{-1},$$

and define $h^{-2} := \ln(\frac{\sigma^d}{\lambda})$, $t_s := |h^{-2} - \sigma^2\|s_{-1}\|^2|$, $\lambda \lesssim \sigma^d \lesssim h^{-d}$. Again we divide the integral into two different domains, $D_1 := \{\sigma^2\|s_{-1}\|^2 < h^{-2}\}$ and $D_2 := \{\sigma^2\|s_{-1}\|^2 \ge h^{-2}\}$. Moreover, we apply residue theorem to the internal integral and similarly have

$$\begin{aligned}
\Big|\int_{-\infty}^{\infty}\frac{e^{2\pi\sqrt{-1}\|u\|s_1}|s_1|^{\mathbf{j}_1}}{1+\frac{\lambda}{\sigma^d}e^{\sigma^2(\|s_{-1}\|^2+s_1^2)}}\mathrm{d}s_1\Big| &= \Big|2\pi\sqrt{-1}\sum_{k=-\infty}^{\infty}\frac{e^{2\pi\sqrt{-1}\|u\|z_{2k-1}}(2\pi\sqrt{-1}z_{2k-1})^{\mathbf{j}_1}}{-2\sigma^2 z_{2k-1}}\Big|\\
&\lesssim \sigma^{-2}\sum_{k=1}^{\infty}e^{2\pi\sqrt{-1}\|u\|z_{2k-1}}(z_{2k-1})^{\mathbf{j}_1-1},
\end{aligned}$$

where $z_{2k-1} = a_k + \sqrt{-1}b_k$, and

$$\begin{aligned} a_k^2 + b_k^2 &= \sigma^{-2}(t_s^2 + ((2k-1)\pi)^2)^{\frac{1}{2}} \\ 2b_k^2 &= (a_k^2 + b_k^2) - (a_k^2 - b_k^2) = \sigma^{-2}\Big[(t_s^2 + ((2k-1)\pi)^2)^{\frac{1}{2}} - t_s\Big] \\ &= \frac{\sigma^{-2}((2k-1)\pi)^2}{(t_s^2 + ((2k-1)\pi)^2)^{\frac{1}{2}} + t_s} \end{aligned}$$

We begin with the first domain D_1, in which $\|s_{-1}\|^2 \le (\sigma h)^{-2}$. We need to set a threshold $H = (\frac{t_s}{\pi} + 1)/2$ for the index k. When $k \le \lfloor H \rfloor$, we have

$$\begin{aligned} 2b_k^2 \ge \sigma^{-2}\frac{(2k-1)^2\pi^2}{3t_s} &\Rightarrow b_k \gtrsim \sigma^{-1}\sqrt{t_s}^{-1}k \\ \sigma^{-2}t_s \le a_k^2 + b_k^2 \le \sqrt{2}\sigma^{-2}t_s &\Rightarrow |z_{2k-1}| = \Theta(\sigma^{-1}\sqrt{t_s}); \end{aligned}$$

when $k \ge \lceil H \rceil$, we have

$$\begin{aligned} 2b_k^2 \ge \sigma^{-2}\frac{(2k-1)^2\pi^2}{3(2k-1)\pi} &\Rightarrow b_k \gtrsim \sigma^{-1}\sqrt{k} \\ \sigma^{-2}(2k-1)\pi \le a_k^2 + b_k^2 \le \sqrt{2}\sigma^{-2}(2k-1)\pi &\Rightarrow |z_{2k-1}| = \Theta(\sigma^{-1}\sqrt{k}); \end{aligned}$$

and the series would be bounded by

$$\begin{aligned} \sigma^{-2}\sum_{k=1}^{\lfloor H \rfloor} e^{-2\pi|u|b_k}|z_{2k-1}|^{\mathbf{j}_1-1} &\lesssim \sigma^{-2}(\sigma^{-1}\sqrt{t_s})^{\mathbf{j}_1-1}\sum_{k=1}^{\lfloor H \rfloor} e^{-C_3|u|\sigma^{-1}\sqrt{t_s}^{-1}k} \\ &\lesssim \sigma^{-(\mathbf{j}_1+1)}\sqrt{t_s}^{\mathbf{j}_1-1}e^{-C_3|u|\sigma^{-1}\sqrt{t_s}^{-1}} \\ \sigma^{-2}\sum_{k=\lceil H \rceil}^{\infty} e^{-2\pi|u|b_k}|z_{2k-1}|^{\mathbf{j}_1-1} &\lesssim \sigma^{-2}\sum_{k=\lceil H \rceil}^{\infty} e^{-C_3|u|\sigma^{-1}\sqrt{k}}(\sigma^{-1}\sqrt{k})^{\mathbf{j}_1-1} \\ &\lesssim \sigma^{-(\mathbf{j}_1+1)}(e^{-C_3|u|\sigma^{-1}\sqrt{\lceil H \rceil}}\sqrt{\lceil H \rceil}^{\mathbf{j}_1-1} + \int_{\lceil H \rceil}^{\infty} e^{-C_3|u|\sigma^{-1}\sqrt{x}}\sqrt{x}^{\mathbf{j}_1-1}\mathrm{d}x) \\ &\lesssim \sigma^{-(\mathbf{j}_1+1)}(e^{-C_3|u|\sigma^{-1}\sqrt{\lceil H \rceil}}\sqrt{\lceil H \rceil}^{\mathbf{j}_1-1} + 1/(C_3|u|\sigma^{-1})e^{-C_3|u|\sigma^{-1}\sqrt{\lceil H \rceil}}\sqrt{\lceil H \rceil}^{\mathbf{j}_1}) \\ &\lesssim \sigma^{-(\mathbf{j}_1+1)}\sqrt{\lceil H \rceil}^{\mathbf{j}_1-1}e^{-C_3|u|\sigma^{-1}\sqrt{\lceil H \rceil}}. \end{aligned}$$

We drop one term in the last line as $\sigma\sqrt{\lceil H \rceil} \le \sigma h^{-1} \lesssim 1$, and note the first series would only appear when $t_s > \pi$, and $\lceil H \rceil = \Theta(\max(t_s/\pi, 1))$.

For the integral over the domain D_1, it would be bounded as

$$\begin{aligned} &\int_{D_1} \frac{e^{2\pi\sqrt{-1}\|u\|s_1}}{1 + \frac{\lambda}{\sigma^d}e^{\sigma^2(\|s_{-1}\|^2 + s_1^2)}}\prod_{i=1}^{d}|s_i|^{\mathbf{j}_i}\mathrm{d}s \le \int_{\|s_{-1}\| \le \frac{1}{\sigma h}}\prod_{i=2}^{d}|s_i|^{\mathbf{j}_i} \cdot \int_{-\infty}^{\infty}\frac{e^{2\pi\sqrt{-1}\|u\|s_1}|s_1|^{\mathbf{j}_1}}{1 + \frac{\lambda}{\sigma^d}e^{\sigma^2\|s_{-1}\|^2}e^{\sigma^2 s_1^2}}\mathrm{d}s_1\mathrm{d}s_{-1} \\ \lesssim &\sigma^{-(\mathbf{j}_1+1)}\int_{\|s_{-1}\| \le \frac{1}{\sigma h}}\|s_{-1}\|^{|\mathbf{j}|-\mathbf{j}_1} \cdot (\sqrt{t_s}^{\mathbf{j}_1-1}e^{-C_3|u|\sigma^{-1}\sqrt{t_s}^{-1}} + \sqrt{\lceil H \rceil}^{\mathbf{j}_1-1}e^{-C_3|u|\sigma^{-1}\sqrt{\lceil H \rceil}})\mathrm{d}s_{-1} \\ \lesssim &\sigma^{-(\mathbf{j}_1+1)}\Big(\int_0^{\frac{1}{\sigma h}} r^{|\mathbf{j}|-\mathbf{j}_1+d-2}\sqrt{t_s}^{\mathbf{j}_1-1}e^{-C_3|u|\sigma^{-1}\sqrt{t_s}^{-1}}\mathrm{d}r + \int_0^{\frac{1}{\sigma h}} r^{|\mathbf{j}|-\mathbf{j}_1+d-2}\sqrt{\lceil H \rceil}^{\mathbf{j}_1-1}e^{-C_3|u|\sigma^{-1}\sqrt{\lceil H \rceil}})\mathrm{d}r\Big). \end{aligned}$$

To bound the first integral term, we should utilize a transformation $r = \frac{\sin(\theta)}{\sigma h}, t_s = h^{-1}\cos(\theta)$, and have

$$\begin{aligned} &\int_0^{\frac{1}{\sigma h}} r^{|\mathbf{j}|-\mathbf{j}_1+d-2}\sqrt{t_s}^{\mathbf{j}_1-1}e^{-C_3|u|\sigma^{-1}\sqrt{t_s}^{-1}}\mathrm{d}r \\ =&(\sigma h)^{-(|\mathbf{j}|-\mathbf{j}_1+d-2)}h^{-(\mathbf{j}_1-1)}\int_0^{\pi/2}\sin(\theta)^{|\mathbf{j}|-\mathbf{j}_1+d-2}\cos(\theta)^{\mathbf{j}_1-1}e^{-C_3|u|\sigma^{-1}h/\cos(\theta)}\mathrm{d}\frac{\sin(\theta)}{\sigma h} \\ \le&(\sigma h)^{-(|\mathbf{j}|-\mathbf{j}_1+d-1)}h^{-(\mathbf{j}_1-1)}\int_0^{\pi/2}e^{-C_3|u|\sigma^{-1}h}\mathrm{d}\theta \lesssim \sigma^{-(|\mathbf{j}|-\mathbf{j}_1+d-1)}h^{-(|\mathbf{j}|+d-2)}e^{-C_3|u|\sigma^{-1}h}; \end{aligned}$$

the second integral term could be addressed by utilizing the fact $\lceil H \rceil \geq 1$

$$\begin{aligned}
&\int_0^{\frac{1}{\sigma h}} r^{|\mathbf{j}|-\mathbf{j}_1+d-2}\sqrt{\lceil H\rceil}^{\mathbf{j}_1-1}e^{-C_3|u|\sigma^{-1}\sqrt{\lceil H\rceil}})\mathrm{d}r \\
\leq &\int_{\sigma^{-1}(h^{-2}-\pi)^{\frac{1}{2}}}^{\frac{1}{\sigma h}} r^{|\mathbf{j}|-\mathbf{j}_1+d-2}e^{-C_3|u|\sigma^{-1}}\mathrm{d}r + \int_0^{\frac{1}{\sigma h}} r^{|\mathbf{j}|-\mathbf{j}_1+d-2}\sqrt{t_s}^{\mathbf{j}_1-1}e^{-C_3|u|\sigma^{-1}}\mathrm{d}r \\
\lesssim &(\sigma h)^{-(|\mathbf{j}|-\mathbf{j}_1+d-1)}h^{-(\mathbf{j}_1-1)}e^{-C_3|u|\sigma^{-1}}\int_0^{\pi/2}\sin(\theta)^{|\mathbf{j}|-\mathbf{j}_1+d-2}\cos(\theta)^{\mathbf{j}_1}\mathrm{d}\theta \\
\leq &(\sigma h)^{-(|\mathbf{j}|-\mathbf{j}_1+d-1)}h^{-(\mathbf{j}_1-1)}e^{-C_3|u|\sigma^{-1}h},
\end{aligned}$$

which could infer the total integral over domain D_1 should be $\mathcal{O}(\sigma^{-(|\mathbf{j}|+d)}h^{-(|\mathbf{j}|+d-2)}e^{-C_3|u|\sigma^{-1}h})$.

Over the second domain D_2, we can similarly divide the series into two parts by the threshold H. We notice $\sqrt{t_s}$ and $\sqrt{k}$ would respectively dominate the scale of b_k or $|z_{2k-1}|$ in the two parts, as

$$\begin{aligned}
2b_k^2 &\geq \sigma^{-2}((2k-1)\pi + t_s) \\
a_k^2 + b_k^2 &= \Theta(\sigma^{-2}((2k-1)\pi + t_s)).
\end{aligned}$$

Using a similar derivation as above, the two parts would be correspondingly bounded by

$$\begin{aligned}
\sigma^{-2}\sum_{k=1}^{\lfloor H\rfloor} e^{-2\pi|u|b_k}|z_{2k-1}|^{\mathbf{j}_1-1} &\lesssim \sigma^{-(\mathbf{j}_1+1)}\sqrt{t_s}^{\mathbf{j}_1-1}\lfloor H\rfloor e^{-C_3|u|\sigma^{-1}\sqrt{t_s}} \\
\sigma^{-2}\sum_{k=\lceil H\rceil}^{\infty} e^{-2\pi|u|b_k}|z_{2k-1}|^{\mathbf{j}_1-1} &\lesssim \sigma^{-(\mathbf{j}_1+1)}\sqrt{\lceil H\rceil}^{\mathbf{j}_1-1}e^{-C_3|u|\sigma^{-1}\sqrt{\lceil H\rceil}}.
\end{aligned}$$

Since $\lceil H\rceil \gtrsim t_s$ and $\lceil H\rceil \geq 1$, the overall series could be bounded by a constant multiple of $\sigma^{-(\mathbf{j}_1+1)}\sqrt{\lceil H\rceil}^{\mathbf{j}_1+1}e^{-C_3|u|\sigma^{-1}\sqrt{\lceil H\rceil}}$. Therefore, by polar coordinate transformation, the integral over D_2 would be reduced to the following one,

$$\begin{aligned}
\int_{\frac{1}{\sigma h}}^{\infty} r^{|\mathbf{j}|-\mathbf{j}_1+d-2}\sqrt{\lceil H\rceil}^{\mathbf{j}_1+1}e^{-C_3|u|\sigma^{-1}\sqrt{\lceil H\rceil}})\mathrm{d}r = &\int_{\frac{1}{\sigma h}}^{\frac{2}{\sigma h}} r^{|\mathbf{j}|-\mathbf{j}_1+d-2}\sqrt{\lceil H\rceil}^{\mathbf{j}_1+1}e^{-C_3|u|\sigma^{-1}\sqrt{\lceil H\rceil}})\mathrm{d}r \\
&+ \int_{\frac{2}{\sigma h}}^{\infty} r^{|\mathbf{j}|-\mathbf{j}_1+d-2}\sqrt{\lceil H\rceil}^{\mathbf{j}_1+1}e^{-C_3|u|\sigma^{-1}\sqrt{\lceil H\rceil}})\mathrm{d}r,
\end{aligned}$$

where we further divide the integral based on whether $t_s > h^{-2}$. For the first stage, utilizing $\lceil H\rceil \geq 1$, we can bound it as

$$\begin{aligned}
\int_{\frac{1}{\sigma h}}^{\frac{2}{\sigma h}} r^{|\mathbf{j}|-\mathbf{j}_1+d-2}\sqrt{\lceil H\rceil}^{\mathbf{j}_1+1}e^{-C_3|u|\sigma^{-1}\sqrt{\lceil H\rceil}})\mathrm{d}r &\lesssim \int_{\frac{1}{\sigma h}}^{\frac{2}{\sigma h}} r^{|\mathbf{j}|-\mathbf{j}_1+d-2}h^{-(\mathbf{j}_1+1)}e^{-C_3|u|\sigma^{-1}}\mathrm{d}r \\
&\lesssim (\sigma h)^{-(|\mathbf{j}|-\mathbf{j}_1+d-1)}h^{-(\mathbf{j}_1+1)}e^{-C_3|u|\sigma^{-1}h};
\end{aligned}$$

for the second stage, we could apply $\lceil H\rceil = \Theta(\sigma^2 r^2)$, and have

$$\begin{aligned}
\int_{\frac{2}{\sigma h}}^{\infty} r^{|\mathbf{j}|-\mathbf{j}_1+d-2}\sqrt{\lceil H\rceil}^{\mathbf{j}_1+1}e^{-C_3|u|\sigma^{-1}\sqrt{\lceil H\rceil}})\mathrm{d}r &\lesssim \int_{\frac{2}{\sigma h}}^{\infty} r^{|\mathbf{j}|-\mathbf{j}_1+d-2}(\sigma r)^{\mathbf{j}_1+1}e^{-C_3|u|r}\mathrm{d}r \\
&\lesssim \sigma^{\mathbf{j}_1+1}(\sigma h)^{-(|\mathbf{j}|+d-1)}e^{-C_3|u|\sigma^{-1}h^{-1}},
\end{aligned}$$

which implies the scale of the integral over D_2 is $\mathcal{O}((\sigma h)^{-(|\mathbf{j}|+d)}e^{-C_3|u|\sigma^{-1}h})$. The second claim can thus be proved by simply combining the current results for D_1 and D_2.

$\diamondsuit$

9 NUMERICAL INTEGRATION

To make the algorithm end in $\widetilde{\mathcal{O}}(n)$ time, we need to efficiently compute all the leverage approximation (6) in the main paper. We first state an observation that the original multiple integral over $\mathbb{R}^d$ could be simplified to a normal integral with only one variable. Then we propose a fast method to give the approximation of the integral, which only requires $\widetilde{\mathcal{O}}(n)$ time to compute all the integrals.

9.1 Simplify the Integration by Polar Coordinate Transformation

An important feature of a Matérn kernel is the isotropy that the value of the kernel function $K_\alpha(x)$ only depends on the module $\|x\|_2$ (for simplicity $\|\cdot\|_2$ would be denoted as $\|\cdot\|$ from then on in this section). The property is shared by the corresponding spectral density $m_\alpha(s)$, and thus the Fourier transform of our rescaled leverage score approximation $\widetilde{K}_\lambda(\cdot,t)$ also inherits the isotropy. In particular, given the center point t and a point x of interest, by Fourier transform formula,

$$\widetilde{K}_\lambda(x,t) = \int_{\mathbb{R}^d} \mathscr{F}[\widetilde{K}_\lambda(\cdot,t)](s)\exp(2\pi\sqrt{-1}x^T s)\mathrm{d}s = \int_{\mathbb{R}^d} \frac{\exp(2\pi\sqrt{-1}(x-t)^T s)}{p(t)+\lambda/m_\alpha(s)}\mathrm{d}s. \tag{7}$$

Considering the specific case $x = t$ in computing leverage score approximation, the value of the integrand would be the same for any s with the same module $\|s\|$. By polar coordinate transformation, we obtain

$$\widetilde{K}_\lambda(t,t) = \int_0^\infty \frac{1}{p(t)+\lambda/m_\alpha(r)} \cdot S_{d-1}(r)\mathrm{d}r$$

where $S_{d-1}(r)$ is the surface area of a (d)-dim ball with radius r.

It is worth mentioning for the general case $x \neq t$, the isotropy could also be utilized to accelerate the computation. In some geostatistics literature, for example, Hankel transform (Kleiber and Nychka, 2015) is applied to simplify the integration into a univariate integral for two-dimensional processes with Matérn kernels. We extend this idea to kernels with dimension more than two and represent the rescaled leverage score approximation $\widetilde{K}_\lambda(x,t)$ as a double integral. We notice the value of the integrand in equation (7) would be the same for any s with the same module $\|s\|$ and the same inner product $(x-t)^T s$. Since the spectral density $m_\alpha(s)$ only depends on $\|s\|$, we slightly abuse the notation, instead representing it as $m_\alpha(\|s\|)$ to emphasize the isotropy. Moreover, by a certain coordinate transformation $r = \|s\|, \cos(\theta) = \frac{(x-t)^T s}{\|x-t\|\|s\|}$, we rewrite the integrand above as $\exp(2\pi\sqrt{-1}\|x-t\|r\cos(\theta))$, and observe that the integrand would remain unchanged with the input points from the intersection between the $(d-1)$-sphere $\{s \in \mathbb{R}^d : \|s\| = r\}$ and the cone $\{s \in \mathbb{R}^d : (x-t)^T s = \|x-t\|\|s\|\cos(\theta)\}$ (the intersection is indeed a $(d-2)$-sphere with radius $r\sin(\theta)$). With those notations, the original d-dim integral would thereby be calculated as

$$\widetilde{K}_\lambda(x,t) = \int_0^\infty \int_0^\pi \frac{\exp(2\pi\sqrt{-1}\|x-t\|r\cos(\theta))}{p(t)+\lambda/m_\alpha(r)} \cdot S_{d-2}(r\sin(\theta))r\mathrm{d}\theta\mathrm{d}r$$

where $S_{d-2}(r\sin(\theta))$ is the surface area of a $(d-1)$-dim ball with radius $r\sin(\theta)$. The same trick applies to all the other stationary kernels with isotropic spectral density function, including Gaussian kernels.

9.2 Approximation of the Integrals

Directly, the integrals above can be computed by a reliable package QUADPACK with the specific integrator QAWF (Piessens et al., 2012), which is targeted at Fourier cosine transform. However, the computation is time-consuming, as QAWF implements an adaptive method so that when $\lambda \to 0$ it will require more function evaluations. To overcome the potential drawback, we propose a fast method to approximate the integration with $o(1)$ relative error in near-constant time.

For Matérn kernels, we would focus on the following integral of a simplified form,

$$\int_0^\infty \frac{x^{d-1}}{p+\lambda(1+x^2)^\alpha}\mathrm{d}x.$$

Inspired by the derivation of the scale of the integral $O(\lambda^{-\frac{d}{2\alpha}})$, we rewrite the integral as

$$\int_0^\infty \frac{x^{d-1}}{p+(\lambda^{\frac{1}{\alpha}}+(\lambda^{\frac{1}{2\alpha}}x)^2)^\alpha}\mathrm{d}x = \lambda^{-\frac{d}{2\alpha}}\int_0^\infty \frac{x^{d-1}}{p+(\lambda^{\frac{1}{\alpha}}+x^2)^\alpha}\mathrm{d}x,$$

and intuitively want to replace $(\lambda^{\frac{1}{\alpha}}+x^2)$ with x^2. We would show the approximation would only result in a small relative error of order $\mathcal{O}(\lambda^{\frac{1}{\alpha}}) = o(1)$ as required.

The difference between the two integrands is

$$\lambda^{-\frac{d}{2\alpha}}\left(\frac{x^{d-1}}{p+x^{2\alpha}} - \frac{x^{d-1}}{p+(\lambda^{\frac{1}{\alpha}}+x^2)^\alpha}\right) = \lambda^{-\frac{1}{2\alpha}}x^{d-1}\frac{(\lambda^{\frac{1}{\alpha}}+x^2)^\alpha - x^{2\alpha}}{(p+x^{2\alpha})(p+(\lambda^{\frac{1}{\alpha}}+x^2)^\alpha)}.$$

When $x^2 \le \lambda^{\frac{1}{\alpha}}$, the numerator above would be bounded by $(2^\alpha-1)\lambda$, and further we have

$$\frac{(\lambda^{\frac{1}{\alpha}}+x^2)^\alpha - x^{2\alpha}}{(p+x^{2\alpha})(p+(\lambda^{\frac{1}{\alpha}}+x^2)^\alpha)} \le \frac{(2^\alpha-1)\lambda}{(p+x^{2\alpha})(p+(\lambda^{\frac{1}{\alpha}}+x^2)^\alpha)} \lesssim \frac{\lambda^{\frac{1}{\alpha}}}{p+x^{2\alpha}}.$$

When $x^2 > \lambda^{\frac{1}{\alpha}}$, we can control the numerator by the first order Taylor approximation,

$$\frac{(\lambda^{\frac{1}{\alpha}}+x^2)^\alpha - x^{2\alpha}}{(p+x^{2\alpha})(p+(\lambda^{\frac{1}{\alpha}}+x^2)^\alpha)} \lesssim \frac{\lambda^{\frac{1}{\alpha}}(x^2)^{\alpha-1}}{(p+x^{2\alpha})(p+(\lambda^{\frac{1}{\alpha}}+x^2)^\alpha)} \lesssim \frac{\lambda^{\frac{1}{\alpha}}}{p+x^{2\alpha}},$$

where the last relation holds as $(x^2)^{\alpha-1} \lesssim p+(\lambda^{\frac{1}{\alpha}}+x^2)^\alpha$.

Then the total difference between two integrals would be bounded by a constant multiple of

$$\lambda^{-\frac{d}{2\alpha}}\int_0^\infty \frac{\lambda^{\frac{1}{\alpha}}x^{d-1}}{p+x^{2\alpha}}\mathrm{d}x,$$

which is $O(\lambda^{-\frac{d}{2\alpha}}\lambda^{\frac{1}{\alpha}})$. Considering the magnitude of the original integral is $\Theta(\lambda^{-\frac{d}{2\alpha}})$, our claim regarding the relative error is validated.

We utilize the formula $\int_0^\infty \frac{dx}{1+x^a} = \frac{\pi/a}{\sin \pi/a}$ to give the final approximation:

$$\int_0^\infty \frac{x^{d-1}}{p+\lambda(1+x^2)^\alpha}\mathrm{d}x \approx p^{\frac{d}{2\alpha}-1}\frac{\lambda^{-\frac{d}{2\alpha}}}{d}\frac{\pi/\frac{d}{2\alpha}}{\sin \pi/\frac{d}{2\alpha}}.$$

As the sampling probability is computed as the normalized leverage, we can even directly use $p^{\frac{d}{2\alpha}-1}$ as the rescaled leverage and ignore the rest factor.

For Gaussian kernels, the formula would be even easier, since there is a closed form expression for the target integral,

$$\frac{2}{\Gamma(d/2)}\int_0^\infty \frac{t^{d-1}}{p(2\pi\sigma^2)^{d/2}+\lambda e^{t^2}}\mathrm{d}t = -\frac{Li_{d/2}(-\frac{p(2\pi\sigma^2)^{d/2}}{\lambda})}{p(2\pi\sigma^2)^{d/2}}$$

where σ is the bandwidth of the Gaussian kernel used, and $Li_{d/2}(\cdot)$ is the polylogarithm function with order $\frac{d}{2}$. The fast computation of the polylogarithm function has already been thoroughly studied by some previous works (Crandall, 2006; Vepštas, 2008; Johansson, 2015), and they proposed various methods to compute $Li_{d/2}(c)$ with $\Theta(\log\log n)$ bits of precision ($\Theta(\frac{1}{\log n})$ relative error) and a polynomial of $\log\log n$ time. The total time to compute the leverage would thus still be $\widetilde{\mathcal{O}}(n)$.

10 DENSITY ESTIMATION

As we described in the main paper, by utilizing the distributional information, leverage scores in a KRR problem could be efficiently approximated by our analytical method. In the case we do not have prior knowledge of

the distribution, we propose to estimate densities of data points via kernel density estimation, which have been discussed in the main paper that by using some recent KDE methods with a sub-optimal error rate, we can perform the density estimation within $\widetilde{\mathcal{O}}(n)$ time. To further justify the usage of kernel density estimation, we imply by the following lemma that given an $o(1)$ error in KDE, the particular error in approximating statistical leverage scores due to the density estimation is asymptotically negligible.

Lemma 12. *Under the same assumptions before, $\forall x \in spt(p)$ (the support of p), given a fixed point t, the supremum of the error caused by the density estimate $\hat{p}(t)$ on point t, is bounded by a constant multiple of $h^{-d}|p(t) - \widehat{p}(t)|$.*

Proof. For simplicity, we first denote the leverage score approximation with estimated density as $\widehat{\widetilde{K}}_\lambda(x, x_i)$. Inserting the density estimation $\hat{p}$ into the formula of rescaled leverage scores (equation 7 in the main paper), we obtain $\mathscr{F}[\widehat{\widetilde{K}}_\lambda(\cdot, x_i)](s) = \frac{e^{-2\pi\sqrt{-1}\langle x_i, s\rangle}}{\widehat{p}(x_i)+\lambda(m_\alpha(s))^{-1}}$.

By triangle inequality, the supremum of the total error $|\widehat{\widetilde{K}}_\lambda(x, x_i) - G(x, x_i)|$ could be divided into two sources, one due to density estimation $|\widehat{\widetilde{K}}_\lambda(x, x_i) - \widetilde{K}_\lambda(x, x_i)|$ and the other due to approximation error $|\widetilde{K}_\lambda(x, x_i) - G_\lambda(x, x_i)|$, which has been thoroughly discussed in the appendix. Here we focus on the first term.

$$\begin{aligned}\sup_{x\in\mathbb{R}^d} |\widehat{\widetilde{K}}_\lambda(x, x_i) - \widetilde{K}_\lambda(x, x_i)| &= \sup_{x\in\mathbb{R}^d} \left|\mathscr{F}^{-1}\left[\mathscr{F}[\widehat{\widetilde{K}}_\lambda(\cdot, x_i)]\right](x) - \mathscr{F}^{-1}\left[\mathscr{F}[\widetilde{K}_\lambda(\cdot, x_i)]\right](x)\right| \\ &= \sup_{x\in\mathbb{R}^d} \left|\mathscr{F}^{-1}\left[\frac{|\widehat{p}(x_i) - p(x_i)|e^{-2\pi\sqrt{-1}\langle x_i, s\rangle}}{(\widehat{p}(x_i) + \lambda\cdot(m_\alpha(s))^{-1})(p(x_i) + \lambda\cdot(m_\alpha(s))^{-1})}\right](x)\right|\end{aligned}$$

$|\widehat{p}(x_i) - p(x_i)|$ could be extracted as a factor, and we solely need to deal with the rest term

$$\begin{aligned}&\sup_{x\in\mathbb{R}^d} \cdot \left|\mathscr{F}^{-1}\left[\frac{e^{-2\pi\sqrt{-1}\langle x_i, s\rangle}}{(\widehat{p}(x_i) + \lambda\cdot(m_\alpha(s))^{-1})(p(x_i) + \lambda\cdot(m_\alpha(s))^{-1})}\right](x)\right| \\ =&\sup_{x\in\mathbb{R}^d} \left|\int_{\mathbb{R}^d} \left[\frac{e^{-2\pi\sqrt{-1}\langle x_i - x, s\rangle}}{(\widehat{p}(x_i) + \lambda\cdot(m_\alpha(s))^{-1})(p(x_i) + \lambda\cdot(m_\alpha(s))^{-1})}\right] ds\right|\end{aligned}$$

Relaxing the exponential term as 1 and using Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}&\left|\int_{\mathbb{R}^d} \left[\frac{e^{-2\pi\sqrt{-1}\langle x_i - x, s\rangle}}{(\widehat{p}(x_i) + \lambda\cdot(m_\alpha(s))^{-1})(p(x_i) + \lambda\cdot(m_\alpha(s))^{-1})}\right] ds\right| \\ \leq&\left|\int_{\mathbb{R}^d} \left[\frac{1}{(\widehat{p}(x_i) + \lambda\cdot(m_\alpha(s))^{-1})(p(x_i) + \lambda\cdot(m_\alpha(s))^{-1})}\right] ds\right| \\ \leq&\left\|\frac{1}{\widehat{p}(x_i) + \lambda\cdot(m_\alpha(s))^{-1}}\right\|_2 \cdot \left\|\frac{1}{p(x_i) + \lambda\cdot(m_\alpha(s))^{-1}}\right\|_2 \\ \lesssim& h^{-d/2}h^{-d/2} = h^{-d}\end{aligned}$$

The last inequality can be verified by Lemma 10. ♢

We finally remark that given the $o(1)$ factor $|\widehat{p}(x_i) - p(x_i)|$, the error $|\widehat{\widetilde{K}}_\lambda(x, x_i) - \widetilde{K}_\lambda(x, x_i)|$ caused by the density estimation would therefore be $o(h^{-d})$, and thus is enough to make the relative error of leverage approximation vanish.

10.1 Missing assumptions for modified HBE

We list some advanced KDE methods in the main paper to show theoretically we can estimate the density with time complexity at most polynomial in the dimension d. Among them, modified Hashing-Based Estimators (HBE) (Backurs et al., 2019) is the most recent one. Taking this method as a representative, we copy the assumption in modified HBE here for the sake of completeness.

Assumption 3 ($(\frac{1}{2}, M)$-LSHable)**.** *Let $\mathcal{K}_e(x, y)$ be the kernel function used for KDE, for which there exists a distribution H of hash functions and $M \geq 1$ such that for every $x, y \in \mathbb{R}^d$,*

$$M^{-1} \cdot \mathcal{K}_e(x, y)^{\frac{1}{2}} \leq \mathbb{P}_{h\sim H}\{h(x) = h(y)\} \leq M \cdot \mathcal{K}_e(x, y)^{\frac{1}{2}}.$$

To attain the fast rate claimed by modified HBE, the core assumption above that the kernel used for KDE is $(\frac{1}{2}, M)$-LSHable for some constant M is necessary. The authors have proved that some common kernels, such as Laplacian and exponential kernels, are $(\frac{1}{2}, \mathcal{O}(1))$-LSHable; and thus a density estimator based on those kernels can be efficiently approximated by modified HBE.

11 TECHNICAL RESULTS

Some tricks in multivariate integrals are heavily utilized in this work, and here we present a lemma to address the technical details about it. We first would like to mention the notation $\int_{\mathbb{R}^d} f(x)\mathrm{d}F_n(x)$ in our paper is not strict in general, as the multivariate version Riemann–Stieltjes integral is not well defined. In this appendix we just abuse the integral $\int_{\mathbb{R}^d} f(x)\mathrm{d}F_n(x)$ to represent the summation $\frac{1}{n}\sum_{i=1}^n f(x_i)$, and the integral $\int_{\mathbb{R}^d} f(x)\mathrm{d}F(x)$ to represent the expectation $\int_{\mathbb{R}^d} f(x)p(x)\mathrm{d}x$.

The lemma is presented as follows. (cf. Section 6.3 for the notations in the lemma.)

Lemma 13 (Multivariate integration by parts)**.** *Given the absolute continuous approximation F with the compact support Ω and L_∞ density p, the empirical distribution F_n, and an integrand $g(\cdot) \in W^{\alpha,2}$ independent of F_n, the certain integral of interest $\int_{C(y,\delta)} g(x)\mathrm{d}(F_n - F)(x)$ is almost surely (considering the samples in F_n are drawn from F) equal to*

$$\sum_{\mathbf{A}\sqcup\mathbf{B}\sqcup\mathbf{C}=[d]} (-1)^{|\mathbf{A}|+|\mathbf{B}|} \int_{C(y_{\mathbf{A}},\delta)} D^{\mathbf{I_A}} g\Big(x_{\mathbf{A}}; \big(y+\delta\cdot(-I_{\mathbf{B}}+I_{\mathbf{C}})\big)_{\mathbf{B}\sqcup\mathbf{C}}\Big)\cdot \Big(F_n - F\Big)\big(x_{\mathbf{A}}; \big(y+\delta\cdot(-I_{\mathbf{B}}+I_{\mathbf{C}})\big)_{\mathbf{B}\sqcup\mathbf{C}}\big)\mathrm{d}x_{\mathbf{A}},$$

where $\sqcup$ is the notation for disjoint union. The sets $\mathbf{B}$ and $\mathbf{C}$ indeed indicate the certain dimensions to which lower and upper limits are assigned separately. Specifically, if $C(x_0,\delta) = \mathbb{R}^d (\delta = \infty)$ and $g(x)$ vanishes at infinity, $\int_{\mathbb{R}^d} g(x)\mathrm{d}(F_n - F)(x) = (-1)^d \int_{\mathbb{R}^d} (F_n - F)(x)\frac{\partial^d g(x)}{\partial x_1 \partial x_2 \cdots \partial x_d}\mathrm{d}x$; if $g(x)$ and its mixed derivative (up to order α) vanish at infinity and are $L-$Lipschitz, the claim would hold without the assumption on the independence between g and F_n.

Proof. Without loss of generality, we would illustrate our claim by a special 2-d case to avoid the tedious calculation. We would first prove the conclusion for an indefinitely differentiable density $p_{n,\epsilon}(x) = \frac{1}{n}\sum_{i=1}^n \eta_\epsilon(x - x_i)$, where the heat kernel $\eta_\epsilon(x) := \frac{1}{\sqrt{2\pi\epsilon}}\exp(-\frac{\langle x,x\rangle}{2\epsilon})$ of the Dirac delta function. (From then on in this proof, x_i means the i-th element of the vector x.) As a sketch of the proof, we would first show the lemma holds for the integral $\int_{C(y,\delta)} g(x)\big(p_{n,\epsilon}(x) - p(x)\big)\mathrm{d}x$, and finally prove as $\epsilon \to 0$, the integral would converge to the claimed expression in this lemma.

For simplicity, we denote $q(x) = p_{n,\epsilon}(x) - p(x)$ and $Q(x)$ is the corresponding distribution function. We further denote $Q_1(x_1; x_2) := \int_{-\infty}^{x_1} q(t, x_2)\mathrm{d}t$ as a 1-d distribution function with a parameter x_2, so that Riemann–Stieltjes integral is applicable to $Q_1(x_1; x_2)$. By definition $\int_{-\infty}^{x_2} Q_1(x_1; t)\mathrm{d}t = Q(x_1, x_2)$, and with that we have

$$\begin{aligned}\int_{C(y,\delta)} g(x)q(x)\mathrm{d}x &= \int_{y_2-\delta}^{y_2+\delta}\int_{y_1-\delta}^{y_1+\delta} g(x_1,x_2)q(x_1,x_2)\mathrm{d}x_1\mathrm{d}x_2 \\ &= \int_{y_2-\delta}^{y_2+\delta}\int_{y_1-\delta}^{y_1+\delta} g(x_1,x_2)\mathrm{d}Q_1(x_1;x_2)\mathrm{d}x_2\end{aligned}$$

We can safely apply integration by parts to the inside integral and obtain:

$$\int_{C(y,\delta)} g(x)q(x)\mathrm{d}x = \int_{y_2-\delta}^{y_2+\delta} \Big(g(x_1,x_2)Q_1(x_1;x_2)\big|_{y_1-\delta}^{y_1+\delta} - \int_{y_1-\delta}^{y_1+\delta} Q_1\frac{\partial g}{\partial x_1}\mathrm{d}x_1\Big)\mathrm{d}x_2 \tag{8}$$

$$= \int_{y_2-\delta}^{y_2+\delta} g(y_1+\delta,x_2)Q_1(y_1+\delta,x_2) - g(y_1-\delta,x_2)Q_1(y_1-\delta,x_2)\mathrm{d}x_2 - \int_{C(y,\delta)} Q_1\frac{\partial g}{\partial x_1}\mathrm{d}x. \tag{9}$$

Now we expand the original integral into three terms. By repeatedly applying integration by parts to the first two terms, we have:

$$\begin{aligned}
&\int_{y_2-\delta}^{y_2+\delta} g(y_1+\delta,x_2)Q_1(y_1+\delta,x_2)\mathrm{d}x_2 = g(y_1+\delta,y_2+\delta)Q(y_1+\delta,y_2+\delta)\\
&\qquad - g(y_1+\delta,y_2-\delta)Q(y_1+\delta,y_2-\delta) - \int_{y_2-\delta}^{y_2+\delta} Q_1(y_1+\delta,x_2)\frac{\partial g(y_1+\delta,x_2)}{\partial x_2}\mathrm{d}x_2\\
&\int_{y_2-\delta}^{y_2+\delta} -g(y_1-\delta,x_2)Q_1(y_1-\delta,x_2)\mathrm{d}x_2 = -g(y_1-\delta,y_2+\delta)Q(y_1-\delta,y_2+\delta)\\
&\qquad + g(y_1-\delta,y_2-\delta)Q(y_1-\delta,y_2-\delta) + \int_{y_2-\delta}^{y_2+\delta} Q_1(y_1+\delta,x_2)\frac{\partial g(y_1-\delta,x_2)}{\partial x_2}\mathrm{d}x_2
\end{aligned}$$

For the last term, we need to change the order of integration and have,

$$\begin{aligned}
&-\int_{C(y,\delta)} Q_1\frac{\partial g}{\partial x_1}\mathrm{d}x\\
=&-\int_{y_1-\delta}^{y_1+\delta}\int_{y_2-\delta}^{y_2+\delta}\frac{\partial g(x_1,x_2)}{\partial x_1}\mathrm{d}Q(x_1,x_2)\mathrm{d}x_1 = -\int_{y_1-\delta}^{y_1+\delta} Q(x_1,y_2+\delta)\frac{\partial g(x_1,y_2+\delta)}{\partial x_1}\mathrm{d}x_1\\
&\quad+\int_{y_1-\delta}^{y_1+\delta} Q(x_1,y_2-\delta)\frac{\partial g(x_1,y_2-\delta)}{\partial x_1}\mathrm{d}x_1 + \int_{C(y,\delta)} Q\frac{\partial^2 g}{\partial x_1\partial x_2}\mathrm{d}x
\end{aligned}$$

Summing up all the nine terms above, we would exactly obtain the claimed equation in the lemma. In particular, if $C(y,\delta)=\mathbb{R}^d(\delta=\infty)$ and $g(x)$ vanishes at infinity, the first two terms in equation (9) would be dropped, and finally the only term left is $\int_{\mathbb{R}^d} g(x)q(x)\mathrm{d}x$, which is equal to $(-1)^d\int_{\mathbb{R}^d} Q(x)\frac{\partial^d g(x)}{\partial x_1\partial x_2\cdots\partial x_d}\mathrm{d}x$ as claimed.

To complete the proof, we still need to show the convergence. We would begin with the assumption $g(\cdot)$ belongs to a dense subset $\mathscr{D}\subset W^{\alpha,2}$, where $\mathscr{D}$ is the space of test functions. Here we simply borrow some definitions and notations from the book (Debnath et al., 2005, Chapter 6), with respect to test functions and weak distributional convergence. A test function is defined as an infinitely differentiable function on $\mathbb{R}^d$ vanishing outside of some bounded set. We denote weak distributional convergence for a sequence of distributions (P_m) to P as $P_m\to P$ if $\langle P_m,g\rangle\to\langle P,g\rangle, \forall g\in\mathscr{D}$. We can see our choice $P_{n,\epsilon}\to F_n$ in the weak distributional sense by setting $P_m=P_{n,1/m}$, and thus for the left hand side of our claim, $\int_{C(y,\delta)} g(x)\mathrm{d}(P_{n,1/m}-F)(x)\to\int_{C(y,\delta)} g(x)\mathrm{d}(F_n-F)(x)$ by some standard techniques; For the terms in the right hand side, we would illustrate by taking $\int_{C(y,\delta)} D^{\mathbf{I}_{[d]}}g(x)(P_{n,\epsilon}-P)(x)\mathrm{d}x$ as an example. We note $D^{\mathbf{I_A}}g$ is still a test function, and $P_{n,\epsilon}$ converges to F_n in L_2, so that by Cauchy-Shwartz inequality we could obtain,

$$\Big|\int_{C(y,\delta)} D^{\mathbf{I}_{[d]}}g(x)(P_{n,\epsilon}-F_n)(x)\mathrm{d}x\Big| \le \Big(\int_{C(y,\delta)}|D^{\mathbf{I}_{[d]}}g(x)|^2\mathrm{d}x\cdot\int_{C(y,\delta)}|(P_{n,\epsilon}-F_n)(x)|^2\mathrm{d}x\Big)^{\frac{1}{2}}\to 0,$$

which implies our desired convergence.

The next step is to extend test functions to functions in $W^{\alpha,2}$. An important fact is that $\mathscr{D}$ is a dense subspace of $W^{\alpha,2}$, and we can find a sequence of test functions converging to g in $W^{\alpha,2}$ and further a subsequence (g_m) converging to g both in $W^{\alpha,2}$ and almost everywhere, since convergence in $W^{\alpha,2}$ implies convergence in L_2. The Cauchy-Shwartz inequality trick for the right hand side would still work as this time the L_2 norm of $D^{\mathbf{I}_{[d]}}\big(g(x)-g_m(x)\big)$ goes to zero. For the left-hand side, as we assume the n samples are drawn from the absolutely continuous distribution F, for the certain sequence above we can show

$$\begin{aligned}
\Big|\int_{C(y,\delta)}\big(g(x)-g_m(x)\big)\mathrm{d}F(x)\Big| &\le \int_{C(y,\delta)}|g(x)-g_m(x)|p(x)\mathrm{d}x\\
&\le \|g(x)-g_m(x)\|_{1,\Omega}\cdot\|p\|_\infty\\
&\le |\Omega|^{\frac{1}{2}}\|p\|_\infty\|g(x)-g_m(x)\|_2\to 0,
\end{aligned}$$

and with probability one, over the n sample points g_m would pointwisely go to g. We can thus conclude our claim would hold for functions in $W^{\alpha,2}$ almost surely.

Finally, for the special case in which $g(x)$ and its mixed derivative (up to order α) are $L-$Lipschitz, we are able to construct a convergent test function sequence g_m which would pointwisely converge to g over the compact support Ω containing all the samples in F_n. We first introduce a sequence of test functions

$$\phi_m := \begin{cases} \frac{m^{\frac{d(d+1)}{2}}}{\varphi} e^{(\|m^{\frac{d+1}{2}}x\|^2-1)^{-1}}, & \text{if } \|x\| < m^{-\frac{d+1}{2}}, \\ 0, & \text{otherwise,} \end{cases}$$

where $\varphi = \int \phi_1(x)\mathrm{d}x$ is the normalization factor. The sequence $\{g_m\}$ is constructed as $\{\phi_m * (g \cdot 1_{m\Omega})\}$, the convolution of the test function ϕ_m and the truncation $g \cdot 1_{m\Omega}$, which is still a sequence of test functions. It can be shown the sequence g_m would go to g in $W^{\alpha,2}$. To validate it, we need to observe the fact that for $x \in m\Omega$,

$$\begin{aligned} |g_m(x) - g(x)| &= |\int \phi_m(t)\big(g(x-t) - g(x)\big)\mathrm{d}t| \le \int_{t<m^{-\frac{d+1}{2}}} \phi_m(t)|g(x-t) - g(x)|\mathrm{d}t \\ &\le \frac{L}{m^{\frac{d+1}{2}}} \int_{t<m^{-\frac{d+1}{2}}} \phi_m(t)\mathrm{d}t = \frac{L}{m^{\frac{d+1}{2}}}, \end{aligned}$$

where the second inequality holds because of the Lipschitz continuity. Therefore,

$$\begin{aligned} \int |g_m(x) - g(x)|^2\mathrm{d}x &= \int_{m\Omega} |g_m(x) - g(x)|^2\mathrm{d}x + \int_{\mathbb{R}^d - m\Omega} |g_m(x) - g(x)|^2\mathrm{d}x \\ &\le \frac{L^2}{m^{d+1}}|m\Omega| + 2\int_{\mathbb{R}^d - m\Omega} g^2(x)\mathrm{d}x + 2\int_{\mathbb{R}^d - m\Omega} g_m^2(x)\mathrm{d}x. \end{aligned}$$

Note the first term is proportional to $1/m$, and $g(x)$ vanishes at infinity. The first two terms would both go to zero as $m \to \infty$. For the last term, we could utilize Jensen's inequality and have $g_m^2(x) \le \int \phi_m(t)g^2(x-t)1_{m\Omega}(x-t)\mathrm{d}t \le \int \phi_m(t)g^2(x-t)\mathrm{d}t$. Then,

$$\begin{aligned} \int_{\mathbb{R}^d - m\Omega} g_m^2(x)\mathrm{d}x &\le \int_{\mathbb{R}^d - m\Omega}\int_{\mathbb{R}^d} \phi_m(t)g^2(x-t)\mathrm{d}t\mathrm{d}x = \int_{\mathbb{R}^d} \phi_m(t)\int_{\mathbb{R}^d - m\Omega} g^2(x-t)\mathrm{d}x\mathrm{d}t \\ &\le \int_{\mathbb{R}^d} \phi_m(t)\int_{(\mathbb{R}^d - m\Omega)+B(m^{-\frac{d+1}{2}})} g^2(x)\mathrm{d}x\mathrm{d}t = \int_{(\mathbb{R}^d - m\Omega)+B(m^{-\frac{d+1}{2}})} g^2(x)\mathrm{d}x \to 0 \end{aligned}$$

where $B(m^{-\frac{d+1}{2}})$ is a ball with radius $m^{-\frac{d+1}{2}}$, and the last convergence holds again since $g(x)$ vanishes at infinity. Combining the pieces above, we can see $g_m \to g$ in L_2 and the similar conclusion holds for all its mixed derivatives up to order α, which means $g_m \to g$ in $W^{\alpha,2}$. In that case, the convergence for the right-hand side of our claim would still hold as in the paragraph above. For the left hand side, this time g_m would uniformly converge to g over Ω, and we can show the integral $\int_{C(y,\delta)} g_m(x)\mathrm{d}(F_n - F)(x)$ would converge to $\int_{C(y,\delta)} g(x)\mathrm{d}(F_n - F)(x)$, even if g depends on the empirical distribution F_n.

$\diamondsuit$

References

Adams, R. A. and Fournier, J. J. (2003). *Sobolev spaces*, volume 140. Elsevier.

Anagnostopoulos, C., Savva, F., and Triantafillou, P. (2018). Scalable aggregation predictive analytics. *Applied Intelligence*, 48(9):2546–2567.

Backurs, A., Indyk, P., and Wagner, T. (2019). Space and time efficient kernel density estimation in high dimensions. In *Advances in Neural Information Processing Systems*, pages 15773–15782.

Brezis, H. and Mironescu, P. (2018). Gagliardo–nirenberg inequalities and non-inequalities: The full story. In *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, volume 35, pages 1355–1376. Elsevier.

Crandall, R. E. (2006). Note on fast polylogarithm computation.

Debnath, L., Mikusinski, P., et al. (2005). *Introduction to Hilbert spaces with applications*. Academic press.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Fukumizu, K. (2008). Elements of positive definite kernel and reproducing kernel hilbert space.

Johansson, F. (2015). Rigorous high-precision computation of the hurwitz zeta function and its derivatives. *Numerical Algorithms*, 69(2):253–270.

Kaya, H. and Tüfekci, P. (2012). Local and global learning methods for predicting power of a combined gas & steam turbine. In *International Conference on Emerging Trends in Computer and Electronics Engineering*.

Kleiber, W. and Nychka, D. W. (2015). Spatial statistics.

Lyon, R. J., Stappers, B., Cooper, S., Brooke, J., and Knowles, J. (2016). Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach. *Monthly Notices of the Royal Astronomical Society*, 459(1):1104–1123.

Musco, C. and Musco, C. (2017). Recursive sampling for the nystrom method. In *Advances in Neural Information Processing Systems*, pages 3833–3845.

Piessens, R., de Doncker-Kapenga, E., Überhuber, C. W., and Kahaner, D. K. (2012). *Quadpack: a subroutine package for automatic integration*, volume 1. Springer Science & Business Media.

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer.

Rudi, A., Calandriello, D., Carratino, L., and Rosasco, L. (2018). On fast leverage score sampling and optimal learning. In *Advances in Neural Information Processing Systems*, pages 5672–5682.

Savva, F., Anagnostopoulos, C., and Triantafillou, P. (2018). Explaining aggregates for exploratory analytics. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 478–487. IEEE.

Tüfekci, P. (2014). Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power & Energy Systems*, 60:126–140.

Vepštas, L. (2008). An efficient algorithm for accelerating the convergence of oscillatory series, useful for computing the polylogarithm and hurwitz zeta functions. *Numerical Algorithms*, 47(3):211–252.