Supplementary material for "Local High-order Regularization on Data Manifolds"

Kwang In Kim, Lancaster University
James Tompkin, Harvard SEAS
Hanspeter Pfister, Harvard SEAS
Christian Theobalt, MPI for Informatics

We introduce the basics of Riemannian geometry, and the corresponding notation and symbol conventions used in the main paper. For a comprehensive introduction, we refer to [Lee, 1997, Jost, 2011, Hein et al., 2007] from which our exposition has been developed. Then, we show the proof of proposition 1 in the main paper and analyze the speed of the convergence of our RKHS norm-based energy estimate to the stabilized regularization energy.

1 Notation and symbol convention

Table 1: Notation and symbol convention used in the main paper.

M	Manifold of dimension m
g	Riemannian metric on M
g_{rs} and g^{rs}	A coordinate representation of g and its inverse
dV	Natural volume element
Δ	The Laplace-Beltrami operator
$\mathcal{X}$	A dataset of points $\{X_1, \cdots, X_u\} \subset \mathbb{R}^n$
X	A point $\in M$ as an element of $\mathcal{X}$
$\mathbf{x}$	A coordinate representation of a point $X \in M$
$N_k(X)$	k-nearest neighbors of X
$\mathcal{N}(X)$	A neighborhood of X determined by a distance on M
$T_X(M)$	A local first-order approximation of M
D	A differential operator
D_0	A differential operator applied to $T_X(M)$
$\|f\|_D$	The norm of f induced by D
$\|f\|_K$	The RKHS norm of f corresponding to a kernel K
φ	Surrogate function as smooth interpolant of $f(N_k(X))$
S	Trace of Hessian of φ
h	φ as a second-order polynomial interpolation
q	φ as a Gaussian kernal interpolation

An m-dimensional *manifold* M is a locally Euclidean[1] topological space of dimension m: At each point $X \in M$, there is an open neighborhood which is homeomorphic[2] to an open subset of $\mathbb{R}^m$. A *chart* on M is a pair (U, ϕ), where U is an open subset of M and $\phi : U \to V$ is a homeomorphism of U onto an open subset V of $\mathbb{R}^m$. For a given point $X \in M$ the components of the map $\phi(X) = (x^1(X), \ldots, x^m(X)) = (x^1, \ldots, x^m)$, are called the *coordinates* of X on U. An *atlas* $\mathcal{A}$ for M is a collection of charts whose domains cover M.

An m-dimensional *smooth manifold* is a topological manifold equipped with a smooth structure: A pair of charts (U, ϕ) and (V, φ) are *smoothly compatible* if either $U \cap V = \emptyset$ or $\phi \cdot \varphi^{-1} : \mathbb{R}^m \to \mathbb{R}^m$ is a C^∞-diffeomorphism.[3] A manifold M is smooth if every pair of members of $\mathcal{A}$ are smoothly compatible.

[1] For rigorous definition, it is also Hausdorff and second countable.

[2] A homeomorphism is a continuous function between topological spaces that has a continuous inverse.

[3] A C^r-diffeomorphism is a r-times continuously differentiable homeomorphism whose inverse is also r-times continuously differentiable.

A tangent vector (field) $\mathbf{v}$ is a linear map $\mathbf{v} : C^\infty(M) \to \mathbb{R}$ that satisfies the derivation rule: For all $f, g \in C^\infty(M)$ and for $X \in M$:

$$\mathbf{v}(fg) = f(X)\mathbf{v}(g) + g(X)\mathbf{v}(f).$$

The set of all tangent vectors $\mathbf{v}_X$ at X is a vector space called the *tangent space* $T_X(M)$. An intuitive interpretation of $T_X(M)$ can be obtained if M is a sub-manifold of a Euclidean space: A tangent vector $\mathbf{v}_X$ is an 'arrow' that is tangent to M at X. Then, the tangent space $T_X(M)$ can be interpreted as a tangent plane at X that is a linear approximation of M at the vicinity of X.

A Riemannian manifold (M, g) is a smooth manifold M provided with a metric g which varies smoothly over $M$and provides an inner product on $T_X(M)$ for all $X \in M$.

2 Proof of proposition 1

In Riemannian normal coordinates ($\mathbf{x} = [x^1, \ldots, x^m]^\top$) centered at each point X, Laplacian evaluation at X becomes the trace of the Hessian matrix given as an evaluation of the Hessian operator at X. Given this, the proof is obtained by slightly modifying the techniques used in the convergence analysis of [Audibert and Tsybakov, 2007] and [Kim et al., 2013]. We include the proof for completeness.

Given an underlying probability distribution P supported by a manifold M, our goal is to show that for each point $X \in M$, $tr[Hh(\mathbf{x})]$ converges to $\Delta f(X)$ as the size $u(t)$ of $\mathcal{X}_{u(t)} = \{X_1, \ldots, X_{u(t)}\} \subset M$ grows.

At each data point $X_i \in \mathcal{X}$, the Hessian $Hf(X_i)$ of f is estimated by fitting h^i to $f|_{\mathcal{N}_\epsilon(X_i)}$, where $\mathcal{N}_\epsilon(X_i) = \mathcal{B}(X_i, \epsilon) \cap \mathcal{X}$, $\mathcal{B}(X_i, \epsilon)$ is the ϵ-neighborhood of X_i in coordinates, i.e., $\mathcal{B}(X, \epsilon) := \{X' : \|\mathbf{x} - \mathbf{x}'\|_{T_X(M)} \leq \epsilon\}$,[4] and $h|_{\mathcal{S}}$ denotes the restriction of a function h on a set $\mathcal{S}$: The Hessian $Hh^i(\mathbf{x}_i)$ of h^i is used as an estimate of $Hf(X_i)$. Since the convergence property is homogeneous, we focus only on a single point $X_i \in M$. For notational convenience, we will omit the index i and furthermore, we will identify a point $X_j \in M$ with its normal coordinate representation $\mathbf{x}_j \in T_{X_i}(M)$ at X_i whenever the latter is defined: The normal coordinate value $\mathbf{x}_j$ of X_j is defined when X_j is included in the *injectivity radius* $\text{inj}(X_i)$ of X_i [Lee, 1997]. Here, we assume that (for sufficiently large u) $\mathcal{N}_\epsilon(X_i) \in \text{inj}(X_i)$. This is possible since in a Riemannian manifold, $\text{inj}(X_i)$ is always positive for any $X_i \in M$. For the points $X_j \notin \text{inj}(X_i)$, the corresponding normal coordinate values are assigned with 0.

Accordingly, $\mathcal{N}_\epsilon(X_i)$ will be represented based on its elements in coordinates $\{\mathbf{g}_1, \ldots, \mathbf{g}_k\}$. Here, we use letter $\mathbf{g}$ instead of $\mathbf{x}$ to stress its indexing within $\mathcal{N}_\epsilon(X_i)$ rather than $\mathcal{X}$. Note that at the normal coordinate chart centered at X_i, the coordinate value $\mathbf{x}_i$ of X_i is zero.

The coefficients of h (Eq. 10 in the main paper):

$$h(\mathbf{x}) = f(0) + \sum_{r=1}^{m} [a]_r x^r + \sum_{r=1, s=r}^{m} [b]_{r,s} x^r x^s, \tag{1}$$

are obtained by solving a weighted least squares problem centered at $\mathbf{x}_i = 0$:

$$\begin{aligned} A \approx B &= \arg\min_Q \|\mathbf{K}(\mathbf{X}Q - \mathbf{f})\|^2 \\ &= (\mathbf{X}^\top \mathbf{K} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{K} \mathbf{f}, \end{aligned} \tag{2}$$

where $\mathbf{X}$ is the design matrix containing the first and the second-order monomials of the coordinate values (centered at $\mathbf{x}_i = 0$) of data points in $\mathcal{X}$:

$$A = \left[\nabla f(0)^\top, \frac{1}{2}\text{vec}[Hf(0)]\right]^\top, B = [\ldots, [a]_r, \ldots, [b]_{r,s}, \ldots]^\top, \mathbf{f} = [f(X_1), \ldots, f(X_u)]^\top, \tag{3}$$

where $\text{vec}(M)$ extracts the upper triangular elements of a symmetric matrix M and forms a vector as a linear alignment of them, and $\mathbf{K}$ is a diagonal weight matrix with $[\mathbf{K}]_{j,j} = K(\mathbf{x}_j, \epsilon)$ and the kernel K is defined as:

$$K(\mathbf{x}, q) = \mathbb{1}_{[\|\mathbf{x}\| < q]}, \tag{4}$$

with $\mathbb{1}_{[S]}$ is the indicator function of the set S. Note that when $X_j \notin \text{inj}(X_i)$, $\mathbf{x}_j = 0$.

[4] For simplicity, we use the ϵ-neighborhood instead of k nearest neighbors $N_k(X_i)$. The convergence in the latter case can easily be established by enforcing $N_k(X_i) \subset \mathcal{B}(X_i, \epsilon)$.

The convergence of $Hf(0)$ is established when $\|A - B\| \to 0$ as $u \to \infty$ and $\epsilon \to 0$.

First, we decompose this deviation as:

$$\|A - B\|^2 \leq \|(\mathbf{X}^\top \mathbf{K}\mathbf{X})^{-1}\|_2 \|\mathbf{K}(\mathbf{X}A - \mathbf{f})\|^2, \tag{5}$$

where we have used $\mathbf{K} = \mathbf{K}^2 = \mathbf{K}^\top$ and $\|C^\top C\|_2^2 = \|C^\top\|_2^2$ for any matrix C. In Eq. 5, the first term depends only on the distribution P on M and it is upper bounded as:

$$\|(\mathbf{X}^\top \mathbf{K}\mathbf{X})^{-1}\|_2 \leq \frac{1}{\|u\epsilon^m \mathcal{E}^{-1} \overline{B} \mathcal{E}^{-1}\|_2} \leq \frac{1}{u\epsilon^{m+4}\lambda_{\overline{B}}}, \tag{6}$$

where

$$\begin{aligned}
\mathcal{E} &= \text{diag}([1/\epsilon, \ldots, 1/\epsilon, 1/\epsilon^2, \ldots, 1/\epsilon^2]^\top), \\
\overline{B} &= \frac{1}{u\epsilon^m} \sum_{j=1}^{u} X(\mathbf{x}_j/\epsilon)^\top X(\mathbf{x}_j/\epsilon) K(\mathbf{x}_j, \epsilon), \\
X(\mathbf{x}) &= [x^1, \ldots, x^m, \ldots, x^r x^s, \ldots,] \in \mathbb{R}^D,
\end{aligned}$$

$D = m + \frac{m(m+1)}{2}$, and $\lambda_{\overline{B}}$ is the smallest eigenvalue of $\overline{B}$. The remainder of this section quantifies $\|\mathbf{K}(\mathbf{X}A - \mathbf{f})\|^2$ and $\lambda_{\overline{B}}$ based on the two regularity assumptions on the Hessian operator H and the probability distribution P.

2.1 Quantifying $\|\mathbf{K}(\mathbf{X}A - \mathbf{f})\|^2$

The deviation between the second-order approximation $\mathbf{K}\mathbf{X}A$ and $\mathbf{K}\mathbf{f}$ depends on the smoothness of f. In particular, we can quantify them based on the boundedness of Hf:

Lemma 1 ([Belward et al., 2008]) *Suppose that the Hessian ($Hf(\mathbf{a}) := H_f(\mathbf{a})$) is Lipschitz continuous with the Lipschitz constant γ. Then*

$$\|\mathbf{K}(\mathbf{X}A - \mathbf{f})\|_2^2 = C_1 \gamma^2 k \epsilon^6 \tag{7}$$

with a constant $C_1 > 0$ where k is the size of $\mathcal{N}_\epsilon(0)$.

Proof: For the simplicity of exposition, let's represent each element of $\mathcal{N}_\epsilon(0) = \{\mathbf{g}_1, \ldots, \mathbf{g}_k\}$ based on its scale and the normalized coordinate values: $\mathbf{g}_j = s_j \mathbf{v}_j$ with $\|\mathbf{v}_j\| = 1$.

Applying the first-order Taylor series remainder formula to f expanded at 0 gives for each point $\mathbf{g}_i$,

$$\begin{aligned}
f(s_j\mathbf{v}_j) - f(0) - s_j \mathbf{v}_j^\top \nabla_f(0) &= \int_0^1 (1-t) s_j \mathbf{v}_j^\top H_f(s_j \mathbf{v}_j t) s_j \mathbf{v}_j dt, \\
\Leftrightarrow f(s_j\mathbf{v}_j) - \frac{1}{2} s_j \mathbf{v}_j^\top H_f(0) s_j \mathbf{v}_j - f(0) - s_j \mathbf{v}_j^\top \nabla_f(0) &= \int_0^1 (1-t) s_j \mathbf{v}_j^\top \left(H_f(s_j \mathbf{v}_j t) - H_f(0)\right) s_j \mathbf{v}_j dt, \quad (8)
\end{aligned}$$

where $\nabla_f := \nabla f$.

Substituting the definition of A (Equation 3) into (8) gives $[\mathbf{K}(\mathbf{X}A - \mathbf{f})]_j = 0$ when $[\mathbf{K}]_{j,j} = 0$ and

$$\begin{aligned}
|[\mathbf{K}(\mathbf{X}A - \mathbf{f})]_j| &= \left| \frac{1}{2} s_j \mathbf{v}_j^\top H_f(0) s_j \mathbf{v}_j - f(s_j \mathbf{v}_j) + f(0) + s_j \mathbf{v}_j^\top \nabla_f(0) \right| \\
&\leq \int_0^1 \left| (1-t) s_j \mathbf{v}_j^\top \left(H_f(0) - H_f(s_j \mathbf{v}_j t)\right) s_j \mathbf{v}_j \right| dt \\
&= \frac{1}{6} \gamma s_j^3, \text{ otherwise.}
\end{aligned}$$

Then

$$\|\mathbf{K}(\mathbf{X}A - \mathbf{f})\|^2 = \sum_{j=1}^{u} [\mathbf{K}(\mathbf{X}A - \mathbf{f})]_j^2 \leq \frac{1}{36} k \gamma^2 \epsilon^6, \tag{9}$$

where we used the fact that only k summands are non-zero and $s_j \leq \epsilon$. □

Substituting Eqs. 9 and 6 into Eq. 5 gives

$$\|A - B\|^2 \leq \frac{1}{36} \frac{1}{\lambda_{\overline{B}}} \frac{k\gamma^2}{u\epsilon^{m-2}}. \tag{10}$$

2.2 Quantifying $\lambda_{\overline{B}}$

Here, we adopt the results of [Audibert and Tsybakov, 2007] to construct a lower bound of $\lambda_{\overline{B}}$. Applying this result requires requires a certain regularity assumption on the underlying probability distribution P on M ([Audibert and Tsybakov, 2007]; Definition 1 of the main paper):
For some constants $c_0, \epsilon_0 > 0$, we will say that a Lebesgue measurable set $A \subset \mathbb{R}^m$ is (c_0, ϵ_0)-*regular* if

$$\lambda[A \cap \mathcal{B}(\mathbf{x}, \epsilon)] \geq c_0 \lambda[\mathcal{B}(\mathbf{x}, \epsilon)], \ \forall \epsilon \in [0, \epsilon_0], \forall \mathbf{x} \in A, \tag{11}$$

where $\lambda[S]$ is the Lebesgue measure of $S \subset \mathbb{R}^m$. We fix constants $c_0, \epsilon_0 > 0$ and $0 < \mu_{\min} < \mu_{\max} < \infty$ and a compact $\mathcal{C} \subset \mathbb{R}^m$. We say that the *strong density assumption* is satisfied if the distribution P is supported on a compact (c_0, ϵ_0)-regular set $A \subseteq \mathcal{C}$ and has a density μ w.r.t. the Lebesgue measure bounded away from zero and infinity on A (between $\mu_{\min}$ and $\mu_{\max}$)

$$\mu_{\min} \leq \mu(\mathbf{x}) \leq \mu_{\max}, \ \forall \mathbf{x} \in A \text{ and } \mu(\mathbf{x}) = 0 \text{ otherwise.} \tag{12}$$

Theorem 1 ([Audibert and Tsybakov, 2007]) *Let P satisfy the strong density assumption. Then, there exist constants $C_2, \mu_0 > 0$ such that for any $0 < \epsilon \leq \epsilon_0$ and any $n \geq 1$,*

$$P^{\otimes u}(\lambda_{\overline{B}} \leq \mu_0) \leq 2D \exp(-C_2 u \epsilon^m), \tag{13}$$

where $P^{\otimes u}$ is the product probability measure according to which the sample is distributed.

Combining Eq. 13 and Eq. 10, we obtain that there are positive constants C_1, C_2, μ_0 with probability larger than $1 - (m^2 + 3m) \exp(-C_2 u \epsilon^m)$,

$$\|A - B\|^2 \leq \frac{C_1}{\mu_0} \frac{k\gamma^2}{u\epsilon^{m-2}}. \tag{14}$$

Adopting the strong density assumption, the probability P_ϵ of sampling a data point from the ϵ-neighborhood of $\mathbf{x}_i = 0$ (which is assumed to be zero) is

$$P_\epsilon = \int_A \mu(\mathbf{x}) \mathbb{1}_{[\|\mathbf{x}\| < \epsilon]} d\mathbf{x} \leq \mu_{\max} \int_A \mathbb{1}_{[\|\mathbf{x}\| < \epsilon]} d\mathbf{x} = \mu_{\max} \upsilon_m \epsilon^m, \tag{15}$$

where $\upsilon_m = \lambda[\mathcal{B}(0, 1)]$ and A is the support of P.

Let's define variables $\{\mathbb{1}_\epsilon(j)\}$

$$\mathbb{1}_\epsilon(j) = \begin{cases} 1 & \text{if } \mathbf{x}_j \in \mathcal{N}_\epsilon(0) \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Applying Hoeffding's inequality to $\{\mathbb{1}_\epsilon(1), \ldots, \mathbb{1}_\epsilon(u)\}$ yields

$$P\left(\sum_{j=1}^{u} \mathbb{1}_\epsilon(j) - uP_\epsilon \geq t\right) \leq \exp\left(-\frac{2t^2}{u}\right). \tag{17}$$

Substituting Eq. 15 into Eq. 17 we obtain

$$P\left(k - (\mu_{\max} \upsilon_m) u \epsilon^m \geq t\right) \leq \exp\left(-\frac{2t^2}{u}\right), \tag{18}$$

which states that $\frac{k}{u\epsilon^d} = \mathcal{O}(1)$.

3 Convergence of the RKHS-norm based energy evaluation to the stabilized energy

In Sec. 4 of the main paper, we use the RKHS-norm based local energy evaluation (Eq. 23 of the main paper)

$$\|q^i\|_D^2 := \sum_{k=1}^{\infty} c_k \int_{T_{X_i}(M)} |D^k q^i|_{\mathbf{x}}|^2 d\mathbf{x} = \|q^i\|_K^2, \tag{19}$$

as an approximation of the stabilized energy (Eq. 21 of the main paper).

This is motivated by their large-scale behaviors: As $u \to \infty$, the kernel parameter σ and the diameter of $N_k(X_i)$ shrink toward zero, and the approximation error (i.e. deviation between the two energies) converges to zero at super-linear speed.

We first note that when we use a Gaussian kernel $K(\mathbf{x}_j, \cdot)$ ($X_j \in N_K(X_i)$; instead of q^i), the series in Eq. 19 converges absolutely. Accordingly, the contribution of all high-order (say, from p to infinite) terms in Eq. 19 is bounded by a monotonically decreasing function G of p [Yuille and Grzywacz, 1988]: Since precisely, $\|K(\mathbf{x}_j, \cdot)\|_D^2 = 1$, G is explicitly given as

$$G(p) = 1 - \sum_{k=1}^{p-1} c_k \int_{T_{X_i}(M)} |D^k q^i|_{\mathbf{x}}|^2 d\mathbf{x}. \tag{20}$$

Furthermore, the corresponding sum of the high-order terms in the stabilized energy is bounded by $[0, G(p)]$. Accordingly, we can determine a p^* so that the finite sum up to the p^*-th summand in Eq. 19 approximates the local energy with any given approximation accuracy level.

Secondly, for any finite order $p' \leq p^*$, the corresponding derivatives of a Gaussian function are given as the p'-th order Hermite polynomials multiplied by the Gaussian. Since a Gaussian function *suppresses* any polynomials, these derivatives decrease rapidly as the corresponding points of evaluation deviates from its center X_i [Kara, 2009]. The speed of this decay is controlled by the width σ^2 of K. Accordingly, for given upper bound s on the approximation error and the integral domain $\mathcal{U}'(X_j)$, σ^2 can be determined such that the deviation between the local energy of $K(\mathbf{x}_j, \cdot)$ (defined based on the integrals over the entire $T_{X_i}(M)$ in Eq. 19) and its restriction to $\mathcal{U}'(X_j)$ becomes smaller than s. This can be shown by straightforwardly evaluating integrals: For instance when $m = 1$ and $(-\mu, \mu) = \mathcal{U}'(0)$, ($X_j$ is assumed to be 0 without loss of generality), the integral of the first-order norm of $K(x) = \exp(-\frac{x^2}{\sigma^2})$ over the entire domain $\mathbb{R} \sim T_0(M)$ is[5]

$$\int_{-\infty}^{\infty} \left| \frac{\partial}{\partial x} \exp\left(-\frac{x^2}{\sigma^2}\right) \right|^2 dx = \int_{-\infty}^{\infty} \frac{4x^2}{\sigma^4} \exp\left(-2\frac{x^2}{\sigma^2}\right) dx = \sqrt{\frac{\pi}{2}} \frac{1}{\sigma}. \tag{21}$$

and the corresponding integral restricted to $\mathcal{U}'(0)$ is obtained as

$$\begin{aligned} \int_{-u}^{u} \left| \frac{\partial}{\partial x} \exp\left(-\frac{x^2}{\sigma^2}\right) \right|^2 dx &= \sqrt{\frac{\pi}{2}} \frac{1}{\sigma} \text{erf}\left(\sqrt{2}\frac{\mu}{\sigma}\right) - \frac{\mu\sqrt{\pi}}{\sigma^2} \text{erf}'\left(\sqrt{2}\frac{u}{\sigma}\right) \\ &= \sqrt{\frac{\pi}{2}} \frac{1}{\sigma} \text{erf}\left(\sqrt{2}\frac{\mu}{\sigma}\right) - \frac{2\mu}{\sigma^2} \exp\left(-2\frac{u^2}{\sigma^2}\right), \end{aligned} \tag{22}$$

where *erf* is the error function of the standard Gaussian distribution. We used the *differentiation under the integral sign* (w.r.t. σ) technique to calculate the integrals in Eqs. 21 and 22. The approximation error of the first-order term is then obtained as,

$$\begin{aligned} & c_1 \left(\int_{-\infty}^{\infty} \left| D^1 \exp\left(-\frac{x^2}{\sigma^2}\right) \right|^2 dx - \int_{-u}^{u} \left| D^1 \exp\left(-\frac{x^2}{\sigma^2}\right) \right|^2 dx \right) \\ &= \frac{\sigma}{2}\sqrt{\frac{\pi}{2}} \left[1 - \text{erf}\left(\sqrt{2}\frac{\mu}{\sigma}\right) - 2\sqrt{\frac{2}{\pi}} \frac{\mu}{\sigma} \exp\left(-2\frac{u^2}{\sigma^2}\right) \right]. \end{aligned} \tag{23}$$

Similarly, the approximation error of the second-order term is given as,

$$\begin{aligned} & c_2 \left(\int_{-\infty}^{\infty} \left| D^2 \exp\left(-\frac{x^2}{\sigma^2}\right) \right|^2 dx - \int_{-u}^{u} \left| D^2 \exp\left(-\frac{x^2}{\sigma^2}\right) \right|^2 dx \right) \\ &= 3\sigma\sqrt{\frac{\pi}{2}} \left[1 - \text{erf}\left(\sqrt{2}\frac{u}{\sigma}\right) + \frac{1}{3}\sqrt{\frac{2}{\pi}} \exp\left(-2\frac{u^2}{\sigma^2}\right) \left(\left(\frac{2u}{\sigma}\right)^3 - \frac{2u}{\sigma} \right) \right]. \end{aligned} \tag{24}$$

[5] Since the Gaussian RKHS energy is shift invariant, we assume that $X_i = 0$ without loss of generality.

For both the first- and the second-order errors, the normalized errors (terms in the outer parentheses in Eqs. 23 and 24) are bounded by $[0, 1]$. When σ is large, the exponential terms in the normalized errors dominates and accordingly, the normalized errors tend to be 1. In this *worse* case, the approximation error decreases linearly with respect to σ decrease. However, when σ is already small, *erf* dominates in the normalized errors[6] and therefore, in this case, the corresponding approximation errors decrease super-linearly. Since iteratively taking the derivatives of a Gaussian with respect to σ^2 yields polynomials, repeatedly applying the differentiation under the integral sign technique shows that all higher-order terms behave exactly the same way: for each k, the approximation error is given as a linear term $\alpha^k\sigma$ (with α^k being a constant) multiplied by the normalized error which is dominated by an error function for small σ. Accordingly, for any order k, the approximation error decreases super-linearly.

Since $D^k q^i$ is a kernel expansion of $N_k(X_i)$, its *effective* support can be limited within a neighborhood that encompass $\{\mathcal{U}'(X_k), \forall X_k \in N_k(X_i)\}$. Then, we can control both σ^2 and the diameter of $N_k(X_i)$ so that the resulting local energy $\|q^i\|_D^2$ well-approximates the integrand in the regularizer (Eq. 7 of the main paper).

$$\|f\|_D^2 := \int_M \sum_{k=1}^{\infty} c_k |D^k f|_X|^2 dV(X), \tag{25}$$

where

$$D^{2k} f = \Delta^k f, D^{2k+1} f = \nabla(\Delta^k f), \tag{26}$$

and where $c_k \geq 0$, $|D^{2k} f|^2 := (D^{2k} f)^2$, and $|D^{2k+1} f|^2 := g(D^{2k+1} f, D^{2k+1} f)$.

References

[Audibert and Tsybakov, 2007] Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633.

[Belward et al., 2008] Belward, J. A., Turner, I. W., and Oqielat, M. N. (2008). Numerical investigations of linear least squares methods for derivative estimation. *ANZIAM Journal*, 50:C844–C857.

[Hein et al., 2007] Hein, M., Audibert, J.-Y., and von Luxburg, U. (2007). Graph Laplacians and their convergence on random neighborhood graphs. *JMLR*, 8:1325–1370.

[Jost, 2011] Jost, J. (2011). *Riemannian Geometry and Geometric Analysis*. Springer, New York, 6th edition.

[Kara, 2009] Kara, M. (2009). An analytical expression for arbitrary derivatives of Gaussian functions $\exp(ax^2)$. *Internal Journal of Physical Sciences*, 4(4):247–249.

[Kim et al., 2013] Kim, K. I., Tompkin, J., and Theobalt, C. (2013). Curvature-aware regularization on Riemannian submanifolds. In *Proc. ICCV*, pages 881–888.

[Lee, 1997] Lee, J. M. (1997). *Riemannian Manifolds- An Introduction to Curvature*. Springer, New York.

[Yuille and Grzywacz, 1988] Yuille, A. L. and Grzywacz, N. M. (1988). The motion coherence theory. In *Proc. ICCV*, pages 344–353.

[6] Note, the exponential term dominates every polynomials in the third summand.