Uniform convergence of adaptive graph-based regularization

Matthias Hein

Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Abstract. The regularization functional induced by the graph Laplacian of a random neighborhood graph based on the data is adaptive in two ways. First it adapts to an underlying manifold structure and second to the density of the data-generating probability measure. We identify in this paper the limit of the regularizer and show uniform convergence over the space of Hölder functions. As an intermediate step we derive upper bounds on the covering numbers of Hölder functions on compact Riemannian manifolds, which are of independent interest for the theoretical analysis of manifold-based learning methods.

1 Introduction

Naturally graphs are inherently discrete objects. However if there exists an underlying continuous structure certain neighborhood graphs can be seen as approximations of the underlying continuous structure. The main goal of this paper is to show how that the smoothness functional $S(f)$ induced by the graph Laplacian of a neighborhood graph built from random samples can be defined in such a way that its continuum limit approximates a desired continuous quantity.
In principle such considerations have been the motivation to build algorithms based on the graph Laplacian for dimensionality reduction, clustering and semi-supervised learning, see e.g. [2, 1, 15, 6]. However the theoretical study of this motivation in particular when the data in $\mathbb{R}^d$ lies on a Euclidean submanifold has been only started quite recently. In [8], see also [3], it was shown that the pointwise limit of the normalized graph Laplacian is the weighted Laplace Beltrami operator. The first work where the limit of $S(f)$ has been studied was [4]. There the limit of $S(f)$ for a single function in the case when the data generating probability measure P has full support in $\mathbb{R}^d$ was derived in a two step process, first $n \to \infty$, then letting the neighborhood size $h \to 0$. In this paper we extend this result in several ways. First we extend it to the setting where the data lies on a submanifold[1] M of $\mathbb{R}^d$, second we introduce data-dependent weights for the graph which are used to control the influence of the density p of P on the limit functional, third we do the limit process $n \to \infty$ and $h \to 0$ simultaneously, so that we actually get rates for $h(n)$ and finally we perform this limit uniformly over the function space of Hölder functions.
We include also an extensive discussion of the properties of the limit smoothness

[1] All the results apply also in the case where P has full d-dimensional support in $\mathbb{R}^d$.

functional and why and how it can be interesting in different learning algorithms such as regression, semi-supervised learning and clustering. In particular the adaptation to the two independent structures inherent to the data, the geometry of the data manifold M and the density p of P, are discussed.

2 Regularization with the graph Laplacian and its continuous limit

The first part of this section introduces the smoothness functional induced by the graph Laplacian for an undirected graph, in particular the neighborhood graph studied in this paper. In the second part we will sketch our main result, the uniform convergence of the smoothness functional induced by the graph Laplacian over the space of α-Hoelder functions. In particular we study the adaptation of the continuous limit functional to the geometry of the data manifold and the density of the data generating measure and possible applications thereof in semi-supervised learning, regression and clustering.

2.1 The graph Laplacian and its induced smoothness functional

Let (V, E) be a undirected graph, where V is the set of vertices with $|V| = n$ and E the set of edges. Since the graph is undirected we have a symmetric adjacency matrix W. Moreover we define the degree function as $d_i = \sum_{j=1}^{n} w_{ij}$. Then it can be shown, see [8], that once one has fixed Hilbert spaces $\mathcal{H}_V, \mathcal{H}_E$ of functions on V and E and a discrete differential operator $\nabla : \mathcal{H}_V \to \mathcal{H}_E$, the graph Laplacian[2] $\Delta : \mathcal{H}_V \to \mathcal{H}_V$ is defined as $\Delta = \nabla^* \nabla$, where ∇^* is the adjoint of d. In the literature one mainly finds two types of graph Laplacian, the normalized one $\Delta_{\text{norm}} = \mathbb{1} - D^{-1}W$ and the unnormalized one $\Delta_{\text{unnorm}} = D - W$, where $D_{ij} = d_i \delta_{ij}$. The smoothness functional $S(f) : \mathcal{H}_V \to \mathbb{R}_+$ induced by the graph Laplacian is defined as

$$S(f) = \langle \nabla f, \nabla f \rangle_{\mathcal{H}_E} = \langle f, \Delta f \rangle_{\mathcal{H}_V} .$$

Note that $S(f)$ defines a semi-norm on $\mathcal{H}_V$. It is can be shown that Δ_{norm} and Δ_{unnorm} induce the same $S(f)$ explicitly given as:

$$S(f) = \frac{1}{2\, n(n-1)} \sum_{i \neq j}^{n} w_{ij} \big(f(i) - f(j)\big)^2 .$$

$S(f)$ coincides for the two graph Laplacians since $S(f)$ is independent of the choice of the inner product in $\mathcal{H}_V$, see [9, sec. 2.1.5]. Note that the smoothness functional $S(f)$ penalizes a discrete version of the first derivative of f.
In this paper we study certain neighborhood graphs that is the weights depend

[2] This holds also for directed graphs, see [9, sec. 2.1].

on the Euclidean distance. The vertex set is an i.i.d. sample $\{X_i\}_{i=1}^n$ of the data generating probability measure P. Of special interest is the case where P has support on a m-dimensional submanifold M in $\mathbb{R}^d$. Similar to Coifman and Lafon in [6] for the continuous case we define the weights of the graph as follows:

$$w_{\lambda,h,n}(X_i, X_j) = \frac{1}{h^m} \frac{k(\|i(X_i) - i(X_j)\|^2 / h^2)}{(d_{h,n}(X_i) d_{h,n}(X_j))^\lambda}, \quad \lambda \in \mathbb{R}.$$

where $d_{h,n}(X_i) = \frac{1}{nh^m} \sum_{j=1}^n k(\|X_i - X_j\|^2 / h^2)$ is the degree function corresponding to the weights k. Note that since k is assumed to have compact support, the parameter h determines the neighborhood of a point. We will denote by $S_{\lambda,h,n}(f)$ the smoothness functional with respect to the weights $w_{\lambda,h,n}$,

$$S_{\lambda,h,n}(f) = \frac{1}{2\, n(n-1)h^2} \sum_{i,j=1}^n (f(X_j) - f(X_i))^2 w_{\lambda,h,n}(X_i, X_j).$$

2.2 The continuous regularizer induced by the weighted Laplacian

The weighted Laplacian is the natural extension of the Laplace-Beltrami operator[3] on a Riemannian manifold, when the manifold is equipped with a measure P which is in our case the probability measure generating the data.

Definition 1 (Weighted Laplacian). *Let (M, g_{ab}) be a Riemannian manifold with measure P where P has a differentiable density p with respect to the natural volume element $dV = \sqrt{\det g}\, dx$, and let Δ_M be the Laplace-Beltrami operator on M. Then we define the s-th weighted Laplacian Δ_s as*

$$\Delta_s := \Delta_M + \frac{s}{p} g^{ab} (\nabla_a p) \nabla_b = \frac{1}{p^s} g^{ab} \nabla_a (p^s \nabla_b) = \frac{1}{p^s} \operatorname{div}(p^s \operatorname{grad}). \qquad (1)$$

The weighted Laplacian induces a smoothness functional $S_{\Delta_s} : C_c^\infty(M) \to \mathbb{R}_+$,

$$S_{\Delta_s}(f) := -\int_M f(\Delta_s f)\, p^s dV = \int_M \langle \nabla f, \nabla f \rangle\, p^s dV,$$

The following sketch of our main Theorem 6 shows that one can choose a function $h(n)$ such that $S_{\lambda,h,n}(f)$ approximates $S_{\Delta_\gamma}(f)$ uniformly for $\gamma = 2 - 2\lambda$.

Sketch of main result *Let $\mathcal{F}_\alpha(s)$ be the ball of radius s in the space of Hölder functions on M. Define $\gamma = 2 - 2\lambda$, then there exists a constant c depending only on k such that for $\alpha \geq 3$ and $h(n) = O(n^{-\frac{\alpha}{2\alpha+2m+m^2+m\alpha}})$,*

$$\sup_{f \in \mathcal{F}_\alpha(s)} \left| S_{\lambda,h,n}(f) - c\, S_{\Delta_\gamma}(f) \right| = O\left(n^{-\frac{\alpha}{2\alpha+2m+m^2+m\alpha}} \right) \quad a.s.$$

[3] The Laplace-Beltrami operator on a manifold M is the natural equivalent of the Laplacian in $\mathbb{R}^d$, defined as

$$\Delta_M f = \operatorname{div}(\operatorname{grad} f) = \nabla^a \nabla_a f$$

We refer to Section 4 for a more detailed account of the results. Let us analyze now the properties of the limit smoothness functional S_{Δ_γ}

$$S_{\Delta_\gamma}(f) = \int_M \|\nabla f\|^2_{T_x M}\, p(x)^{2-2\lambda} \sqrt{\det g}\, dx$$

Note first that $\|\nabla f\|_{T_x M}$ is the norm of the gradient of f on M. The meaning becomes clearer when we express $\|f\|_{T_x M}$ as a local Lipschitz constant[4] $L^M_x(f)$,

$$\|\nabla f\|_{T_x M} = L^M_x(f) = \sup_{y \in M} \frac{|f(x) - f(y)|}{d_M(x,y)} \neq \sup_{y \in M} \frac{|f(x) - f(y)|}{\|x - y\|} = L^{\mathbb{R}^d}_x(f)$$

Most important the smoothness of f is measured with respect to the metric of M or in other words with respect to the intrinsic parameterization. That is a small $\|f\|_{T_x M}$ implies that $f(x) \simeq f(y)$ if x and y are close in the metric of M but not in the metric[5] of $\mathbb{R}^d$. Therefore as desired the graph Laplacian based smoothness functional adapts to the intrinsic geometry of the data.
Next we motivate how the adaptation to the density p controlled by λ can be used in learning algorithms. For $\gamma > 0$ the functional S_{Δ_γ} prefers functions f which are smooth in high-density regions whereas changes are less penalized in low-density regions. This is a desired property for semi-supervised learning where one assumes especially if one has only a few labeled points that the classifier should be almost constant in high-density regions whereas changes of the classifier are allowed in low-density regions, see e.g [4]. However also the case $\gamma < 0$ is interesting. Then minimizing $S_{\Delta_\gamma}(f)$ implies the opposite: smoothness of the function f is enforced where one has little data, and more variation of f is allowed where more data points are sampled. Such a penalization seems appropriate for regression and has been considered by Canu and Elisseeff in [5]. Another application is spectral clustering. The eigenfunctions of Δ_γ can be seen as the limit partioning of spectral clustering for the normalized graph Laplacian (however a rigorous proof has not been given yet). We show now that for $\gamma > 0$ the eigenfunction correponding to the first non-zero eigenvalue is likely to change its sign in a low-density region. Let us assume for a moment that M is compact without boundary and that $p(x) > 0, \forall\, x \in M$, then the eigenspace for the first eigenvalue $\lambda_0 = 0$ is given by the constant functions. The next eigenvalue λ_1 can be determined by the Rayleigh-Ritz variational principle

$$\lambda_1 = \inf_{u \in C^\infty(M)} \left\{ \frac{\int_M \|\nabla u\|^2 p(x)^\gamma dV(x)}{\int_M u^2(x) p(x)^\gamma dV(x)} \;\middle|\; \int_M u(x)\, p(x)^\gamma dV(x) = 0 \right\}.$$

Since the first eigenfunction has to be orthogonal to the constant functions, it has to change its sign. However since $\|\nabla u\|^2$ is weighted by p^γ it is obvious that for $\gamma > 0$ the function changes its sign in a region of low density.

[4] f is continuously differentiable so both terms coincide.
[5] Note that small $\|x - y\|$ does not imply that $d_M(x,y)$ is small e.g. imagine a spiral.

3 Covering numbers for α-Hölder functions on compact Riemannian manifolds with boundary

In this section we derive bounds on the covering numbers of α-Hölder functions on compact Riemannian manifolds with boundary. This generalizes the classical bounds for Euclidean space derived by Kolmogorov and Tihomirov [12, 14]. We use in this section the following short notation: For any vector $k = (k_1, \ldots, k_d)$ of d integers, $D^k = \frac{\partial^k}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}}$ with $k = \sum_{i=1}^d k_i$.

3.1 Technicalities on compact Riemannian manifolds with boundary

We briefly introduce the framework of manifolds with boundary of bounded geometry developed by Schick in [13] for non-compact Riemannian manifolds, which we will use frequently in the following. It makes explicit several geometric properties which are usually implicitly assumed due to compactness of the manifold. Note that the boundary ∂M is an isometric submanifold of M. It has a second fundamental form $\overline{I\!I}$ which should not be mixed up with the second fundamental form $I\!I$ of M with respect to the ambient space $\mathbb{R}^d$. We denote by $\overline{\nabla}$ the connection and by $\overline{R}$ the curvature of ∂M. Moreover let ν be the normal inward vector field at ∂M and let K be the normal geodesic flow defined as $K : \partial M \times [0, \infty) \to M : (x', t) \to \exp_{x'}^M(t\nu_{x'})$. Then the collar set $N(s)$ is defined as $N(s) := K(\partial M \times [0, s])$ for $s \geq 0$.

Definition 2 (Manifold with boundary of bounded geometry). *Let M be a manifold with boundary ∂M (possibly empty). It has bounded geometry if the following holds:*

- *(N) Normal Collar: there exists $r_C > 0$ so that the geodesic collar*

$$\partial M \times [0, r_C) \to M : (t, x) \to \exp_x(t\nu_x)$$

  is a diffeomorphism onto its image (ν_x is the inward normal vector).
- *(IC) The injectivity radius*[6] *$r_{\text{inj}}(\partial M)$ of ∂M is positive.*
- *(I) Injectivity radius of M: There is $r_i > 0$ so that if $r \leq r_i$ then for $x \in M \backslash N(r)$ the exponential map is a diffeomorphism on $B_M(0, r) \subset T_x M$ so that normal coordinates are defined on every ball $B_M(x, r)$ for $x \in M \backslash N(r)$.*
- *(B) Curvature bounds: For every $k \in \mathbb{N}$ there is C_k so that $|\nabla^i R| \leq C_k$ and $\overline{\nabla}^i \overline{I\!I} \leq C_k$ for $0 \leq i \leq k$.*

The injectivity radius makes no sense at the boundary since $\text{inj}(x) \to 0$ as $d(x, \partial M) \to 0$. Therefore we replace next to the boundary normal coordinates with normal collar coordinates. In our proofs we divide M into the set $N(r)$[7] and $M \backslash N(r)$. On $M \backslash N(r)$ we work like on a manifold without boundary and on $N(r)$ we use normal collar coordinates defined below.

[6] The injectivity radius $\text{inj}(x)$ at a point x is the largest r such that the exponential map $\exp_x$ is defined on $B_{\mathbb{R}^m}(0, r)$ and injective. In general we refer the reader to Section 2.2. of [9] for basic notions of differential geometry needed in this paper.

[7] Note that for sufficiently small r, $N(r) = \{x \in M \mid d(x, \partial M) \leq r\}$.

Definition 3 (normal collar coordinates). *Let M be a Riemannian manifold with boundary ∂M. Fix $x' \in \partial M$ and an orthonormal basis of $T_{x'}\partial M$ to identify $T_{x'}\partial M$ with $\mathbb{R}^{m-1}$. For $r_1, r_2 > 0$ sufficiently small (such that the following map is injective) define normal collar coordinates,*

$$n_{x'} : B_{\mathbb{R}^{m-1}}(0, r_1) \times [0, r_2] \to M : (v, t) \to \exp^M_{\exp^{\partial M}_{x'}(v)}(t\nu).$$

The tuple (r_1, r_2) is called the width of the normal collar chart $n_{x'}$ and we denote by $n(x', r_1, r_2)$ the set $n_{x'}(B_{\mathbb{R}^{m-1}}(0, r_1) \times [0, r_2])$.

We denote further by $n(x, r)$ the set $\exp_x(B_{\mathbb{R}^m}(0, r))$. The next lemma is often used in the following.

Lemma 1 ([13]). *Let (M, g) be a m-dimensional Riemannian manifold with boundary of bounded geometry. Then there exists $R_0 > 0$ and constants $S_1 > 0$ and S_2 such that for all $x \in M$ and $r \leq R_0$ one has*

$$S_1 r^m \leq \mathrm{vol}(B_M(x, r)) \leq S_2 r^m, \quad \forall x \in M$$
$$S_1 r^m \leq \mathrm{vol}(n(x', r, r)) \leq S_2 r^m, \quad \forall x' \in \partial M$$

Definition 4 (radius of curvature). *The radius of curvature of M is defined as $\rho = \frac{1}{\Pi_{\max} + \Pi_{\max}}$, where $\Pi_{\max} = \sup_{x \in M} \|\Pi\|_x$ and $\overline{\Pi}_{\max} = \sup_{x \in \partial M} \left\|\overline{\Pi}\right\|_x$.*

The radius of curvature tells us how much the manifold M and its boundary ∂M are curved with respect to the ambient space $\mathbb{R}^d$. It is used in the next lemma to compare distances in $\mathbb{R}^d$ with distances in M.

Lemma 2 ([9]). *Let M have a finite radius of curvature $\rho > 0$. We further assume that $\kappa := \inf_{x \in M} \inf_{y \in M \setminus B_M(x, \pi\rho)} \|x - y\| > 0$. Then $B_{\mathbb{R}^d}(x, \kappa/2) \cap M \subset B_M(x, \kappa) \subset B_M(x, \pi\rho)$. Particularly, if $x, y \in M$ and $\|x - y\| \leq \kappa/2$,*

$$\frac{1}{2} d_M(x, y) \leq \|x - y\|_{\mathbb{R}^d} \leq d_M(x, y) \leq \kappa.$$

Note that for a compact manifold (with boundary) one has $\rho > 0$ and $\kappa > 0$.

3.2 Covering numbers for α-Hölder functions

Definition 5 (α-Hölder functions). *For $\alpha > 0$ denote by $\underline{\alpha}$ the greatest integer smaller than α. Let M be a compact Riemannian manifold and let $(U_i, \phi_i)_{i \in I}$ be an atlas of normal coordinate charts, $\phi_i : U_i \subset \mathbb{R}^m \to M$, such that $M \subset \cup_i \phi_i(U_i)$. Then for a $C^{\underline{\alpha}}$-function $f : M \to \mathbb{R}$, let*

$$\begin{aligned} \|f\|_\alpha = &\max_{k \leq \underline{\alpha}} \sup_{i \in I} \sup_{x \in U_i} \left| D^k (f \circ \phi_i)(x) \right| \\ &+ \max_{k = \underline{\alpha}} \sup_{i \in I} \sup_{x, y \in U_i} \frac{\left| D^k (f \circ \phi_i)(x) - D^k (f \circ \phi_i)(y) \right|}{d(x, y)^{\alpha - \underline{\alpha}}} \end{aligned}$$

The function space $\mathcal{F}_\alpha = \{f \in C^{\underline{\alpha}} \mid \|f\|_\alpha < \infty\}$ is the Banach space of Hölder functions. $\mathcal{F}_\alpha(s)$ denotes a ball of radius s in $\mathcal{F}_\alpha$. We define further

$$\|f\|_{C^k(M)} = \max_{k \leq \underline{\alpha}} \sup_{i \in I} \sup_{x \in U_i} \left| D^k (f \circ \phi_i)(x) \right|.$$

Note that since all transition maps between normal charts and their derivatives are uniformly bounded the above definition of $\|f\|_{C^k(M)}$ could be equivalently replaced[8] with the invariant (coordinate independent) norm of the k-th derivatives defined by Hebey in [7] as, $\left\|\nabla^k f\right\|^2 = g^{i_1 j_1} \dots g^{i_k j_k} \nabla_{i_1} \dots \nabla_{i_k} f \, \nabla_{j_1} \dots \nabla_{j_k} f$. For the Lipschitz type condition it is unclear if there exists an equivalent invariant definition. However the following results for $\mathcal{F}_\alpha$ remain true if we assume that the $\underline{\alpha} + 1$-first derivatives are uniformly bounded. This small change leads for sure to a norm which is equivalent to a coordinate independent norm on M.
In order to construct a covering of $\mathcal{F}_\alpha$ we first need a covering of M with normal and normal collar charts.

Theorem 1. *Let M be a compact m-dimensional Riemannian manifold and let $\epsilon \leq \min\{R_0, \operatorname{inj}(\partial M), r_i\}$. Then there exists a maximal ϵ-separated subset $T_1 := \{x'_{i_1}\}_{i \in I_1}$ of ∂M and a maximal ϵ-separated subset $T_2 := \{x_{i_2}\}_{i \in I_2}$ of $M \backslash N(\epsilon)$ such that*

- $N(\epsilon) \subset \bigcup_{i \in I_1} n(x'_i, \epsilon, \epsilon)$ *and* $M \backslash N(\epsilon) \subset \bigcup_{i \in I_2} n(x_i, \epsilon)$,
- $|I_1| \leq 2 \frac{S_2 \operatorname{vol}(\partial M)}{S_1} \left(\frac{2}{\epsilon}\right)^{m-1}$, *and* $|I_2| \leq \frac{\operatorname{vol}(M)}{S_1} \left(\frac{2}{\epsilon}\right)^m$.

Theorem 2. *Let M be a compact m-dimensional manifold and let $s > 0$ and $\epsilon \leq (3se^{2m})(\min\{R_0, \operatorname{inj}(\partial M), r_i\})^\alpha$. Then there exists a constant K depending only on α, m and M such that*

$$\log \mathcal{N}(\epsilon, \mathcal{F}_\alpha(s), \|\cdot\|_\infty) \leq K \left(\frac{s}{\epsilon}\right)^{\frac{m}{\alpha}}$$

The proof of these theorems can be found in the appendix. The main differences of the proof of Theorem 2 to the classical one in [12, 14] are that the function and its derivatives are discretized in different normal charts so that one has to check that coordinate changes between these normal charts do not destroy the usual argument and an explicit treatment of the boundary.

4 Uniform convergence of the smoothness functional induced by the graph Laplacian

4.1 Assumptions

We ignore in this paper measurability problems, see [14] for a discussion. All results in this section are formulated under the following assumptions on the submanifold M, the density p and the kernel k:

Assumption 1 – *$i : M \to \mathbb{R}^d$ is a smooth, isometric embedding,*
– *M is a smooth compact manifold with boundary (∂M can be empty),*
– *P has a density p with respect to the natural volume element dV on M,*
– *$p \in C^3(M)$ and $p(x) > 0, \forall\, x \in M$,*

[8] in the sense that the resulting norms are equivalent

– *the sample* $X_i, i = 1, \ldots, n$ *is drawn i.i.d. from* P.
– $k : \mathbb{R}_+^* \to \mathbb{R}$ *is measurable, non-negative and non-increasing,*
– $k \in C^2(\mathbb{R}_+^*)$, *that is in particular* k, $\frac{\partial k}{\partial x}$ *and* $\frac{\partial^2 k}{\partial x^2}$ *are bounded,*
– k *has compact support on* $[0, R_k^2]$,
– $k(0) = 0$, *and* $\exists r_k > 0$ *such that* $k(x) \geq \frac{\|k\|_\infty}{2}$ *for* $x \in]0, r_k]$.

Since M is compact, M is automatically a manifold of bounded geometry. In particular all curvatures (intrinsic as well as extrinsic) are bounded. In order to emphasize the distinction between extrinsic and intrinsic properties of the manifold we always use the slightly cumbersome notations $x \in M$ (intrinsic) and $i(x) \in \mathbb{R}^d$ (extrinsic). The kernel functions k which are used to define the weights of the graph are always functions of the squared norm in $\mathbb{R}^d$. The condition $k(0) = 0$ implies that the graph has no loops[9]. In particular the kernel is not continuous at the origin. All statements could also be proved without this condition. The advantage of this condition is that some estimators become thereby unbiased. Finally let us introduce the notation , $k_h(t) = \frac{1}{h^m} k\left(\frac{t}{h^2}\right)$. and the following two constants related to the kernel function k,

$$C_1 = \int_{\mathbb{R}^m} k(\|y\|^2) dy < \infty, \quad C_2 = \int_{\mathbb{R}^m} k(\|y\|^2) y_1^2 dy < \infty. \tag{2}$$

4.2 Results and Proofs

The smoothness functional $S_{\lambda,h,n}(f)$ has been defined in Section 2 as

$$S_{\lambda,h,n}(f) = \frac{1}{2\, n(n-1) h^2} \sum_{i,j=1}^{n} (f(X_j) - f(X_i))^2 \frac{1}{h^m} \frac{k(\|i(X_i) - i(X_j)\|^2 / h^2)}{(d_{h,n}(X_i) d_{h,n}(X_j))^\lambda}.$$

Note that this sum is a U-statistic of order 2. We define further p_h as the convolution of the density with the kernel

$$p_h(x) = \int_M k_h(\|i(x) - i(y)\|^2) p(y) \sqrt{\det g}\, dy. \tag{3}$$

and $\tilde{S}_{\lambda,h,n}(f)$ as $S_{\lambda,h,n}(f)$ with $d_{h,n}(x)$ replaced by $p_h(x)$. The following proposition will be often used.

Proposition 1 ([8]). *For any* $x \in M \backslash \partial M$, *there exists an* $h_0(x) > 0$ *such that for all* $h < h_0(x)$ *and any* $f \in C^3(M)$,

$$\begin{aligned} &\int_M k_h \left(\|i(x) - i(y)\|_{\mathbb{R}^d}^2 \right) f(y) p(y) \sqrt{\det g}\, dy \\ &= C_1 p(x) f(x) + \frac{h^2}{2} C_2 \Big(p(x) f(x) S(x) + (\Delta_M(pf))(x) \Big) + O(h^3), \end{aligned}$$

where $S(x) = \frac{1}{2}\Big[-R\big|_x + \frac{1}{2} \|\sum_a I\!I(\partial_a, \partial_a)\|^2_{T_{i(x)}\mathbb{R}^d} \Big]$ *and* $O(h^3)$ *is a function depending on* x, $\|f\|_{C^3}$ *and* $\|p\|_{C^3}$.

[9] An edge from a vertex to itself is called a loop.

Furthermore we use the following result which basically identifies the extended degree-function of the graph defined as $d_{h,n}(x) = \frac{1}{n}\sum_{i=1}^{n} k_h(\|x - X_i\|)$, as a kernel density estimator on the submanifold M.

Proposition 2 (Pointwise consistency of $d_{h,n}(x)$ [9]). *Let $x \in M/\partial M$, then there exist constants b_1, b_2 such that*

$$\mathrm{P}(|d_{h,n}(x) - p_h(x)| > \epsilon) \leq 2\exp\Big(-\frac{nh^m\,\epsilon^2}{2b_2 + 2/3b_1\,\epsilon}\Big)$$

In particular if $h \to 0$ and $nh^m/\log n \to \infty$, $\lim_{n\to\infty} d_{h,n}(x) = C_1\,p(x)$ a.s..

We refer to [9] for a comparison with a similar result of Hendricks et al. in [10]. In [9] the limit of the smoothness functional $S_{\lambda,h,n}$ was shown for a single function using a Bernstein-type inequality of Hoeffding [11] for U-statistics.

Theorem 3 (Strong consistency of the smoothness functional $S_{\lambda,h,n}$). *Let $f \in C^3(M)$. If $h \to 0$ and $nh^m/\log n \to \infty$,*

$$\lim_{n\to\infty} S_{\lambda,h,n}(f) = \frac{C_2}{2C_1^{\lambda}} \int_M \|\nabla f\|^2_{T_xM}\, p(x)^{2-2\lambda}\sqrt{\det g}\; dx, \quad \text{almost surely.}$$

We extend now this theorem to uniform convergence over balls in the function space of α-Hölder functions. As a first step we prove an abstract uniform convergence result without specifying the function class $\mathcal{F}$.

Theorem 4. *Let $\mathcal{F}$ be a function class with $\sup_{f\in F}\sup_{x\in M}\|\nabla_x f\| \leq s$. Then there exist constants $C', C > 0$ such that for all $\frac{C's^2}{nh^m} < \epsilon < 1/C$ and $0 < h < h_{\max}$, with probability greater than $1 - 2\Big(C\,n + \mathcal{N}\big(\frac{\epsilon\,h^{m+1}}{2C\,s}, \mathcal{F}, \|\cdot\|_{\infty}\big)\Big)e^{-\frac{nh^m\,(1/s)^4\,\epsilon^2}{4C}}$,*

$$\sup_{f\in\mathcal{F}} \big|S_{\lambda,h,n}(f) - \mathbb{E}\,\tilde{S}_{\lambda,h,n}(f)\big| \leq \epsilon$$

Proof: First we decompose the term as follows:

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} |S_{\lambda,h,n}(f) - \mathbb{E}\,\tilde{S}_{\lambda,h,n}(f)| \\
&\leq \sup_{f\in\mathcal{F}} |S_{\lambda,h,n}(f) - \tilde{S}_{\lambda,h,n}(f)| + \sup_{f\in\mathcal{F}} |\tilde{S}_{\lambda,h,n}(f) - \mathbb{E}\,\tilde{S}_{\lambda,h,n}(f)| =: I + II
\end{aligned}$$

We start with the term I. Define $U_{n,h}(f) = \frac{2R_k^2\,s^2}{n(n-1)}\sum_{i,j=1}^{n} k_h(\|i(X_i) - i(X_j)\|)$ and let us work in the following on the event $\mathcal{E}_1$ where

$$\max_{1\leq i\leq n} |d_{h,n}(X_i) - p_h(X_i)| \leq \tau, \quad \text{and} \quad |U_{n,h}(f) - \mathbb{E}\,U_{n,h}(f)| \leq \tau$$

From Proposition 2 and the proof of Theorem 3 we know that there exists a constant C such that $\mathcal{E}_1$ holds with probability greater than $1 - Cne^{-\frac{nh^m\tau^2}{C}}$ for

$\tau \geq \frac{C}{nh^m}$. Since M is compact, we have $\forall x \in M$, $0 < D_1 \leq p_h(x) \leq D_2$. Using a Taylor expansion of $x \to x^{-\lambda}$ with

$$\beta = \min\{d_{h,n}(X_i)d_{h,n}(X_j), p_h(X_i)p_h(X_j)\}^{-\lambda-1} \leq (D_1 - \tau)^{-2(\lambda+1)},$$

we get for $\tau < D_1/2$,

$$\left| \frac{1}{(d_{h,n}(X_i)d_{h,n}(X_j))^{\lambda}} - \frac{1}{(p_h(X_i)p_h(X_j))^{\lambda}} \right| \leq \lambda\, \beta\, [(D_2 + \tau)\tau + D_2\tau] \leq C'\tau,$$

where C' is independent of X_i and X_j. By Lemma 1 and 2 we get for $hR_k \leq \min\{\kappa/2, R_0/2\}$, $\mathbb{E}\, U_{n,h}(f) \leq 2^{m+1} S_2\, R_k^{m+2}\, s^2\, D_2\, \|k\|_\infty$ so that for $\tau \leq \mathbb{E}\, U_{n,h}(f)$ we get on $\mathcal{E}_1$

$$\begin{aligned}\sup_{f\in\mathcal{F}} |S_{\lambda,h,n}(f) - \tilde{S}_{\lambda,h,n}(f)| &\leq \frac{2R_k^2\, s^2\, C'\, \tau}{n(n-1)} \sum_{i,j=1}^{n} k_h(\|i(X_i) - i(X_j)\|) \\ &\leq \big(2^{m+1} S_2\, R_k^{m+2}\, s^2\, D_2\, \|k\|_\infty + \tau\big) C'\tau \leq \frac{\epsilon}{4},\end{aligned}$$

where we have set $\tau = \frac{\epsilon}{C'\, 2^{m+2} S_2\, R_k^{m+2}\, s^2\, D_2\, \|k\|_\infty}$. Now let us deal with II. By assumption we have a δ-covering of $\mathcal{F}$ in the $\|\cdot\|_\infty$-norm. We rewrite the U-statistic $\tilde{S}_{\lambda,h,n}(f) = \frac{1}{n(n-1)} \sum_{i,j} h_f(X_i, X_j)$ with kernels h_f indexed by $f \in \mathcal{F}$, where

$$h_f(x,y) = \frac{1}{h^{m+2}} \frac{k(\|i(x) - i(y)\|)}{(p_h(x)p_h(y))^{\lambda}} \big[f(x) - f(y)\big]^2$$

The δ-covering $C_\delta(\mathcal{F})$ of $\mathcal{F}$ induces a covering of $\mathcal{H}_{\mathcal{F}} = \{h_f \,|\, f \in \mathcal{F}\}$.

$$|h_f(x,y) - h_g(x,y)| \leq \frac{8\, \|k\|_\infty}{h^{1+m} D_1^{2\lambda}}\, s\, R_k\, \|f - g\|_\infty \leq \frac{C}{2}\, s\, \frac{\|f - g\|_\infty}{h^{m+1}},$$

where we have used Lemmas 1,2 and have set $C = \frac{16\|k\|_\infty R_k}{D_1^{2\lambda}}$. We conclude that a δ-covering of $\mathcal{F}$ induces a $\frac{Cs}{h^{m+1}}\delta$-covering of $\mathcal{H}_{\mathcal{F}}$. This implies that for any $f \in \mathcal{F}$ there exists a $g \in C_\delta(\mathcal{F})$ such that,

$$|\tilde{S}_{\lambda,h,n}(f) - \mathbb{E}\, \tilde{S}_{\lambda,h,n}(f)| \leq C\frac{s\,\delta}{h^{m+1}} + |\tilde{S}_{\lambda,h,n}(g) - \mathbb{E}\, \tilde{S}_{\lambda,h,n}(g)|$$

We denote by $\mathcal{E}_2$ the event where $\sup_{g\in C_\delta(\mathcal{F})} |\tilde{S}_{\lambda,h,n}(g) - \mathbb{E}\, \tilde{S}_{\lambda,h,n}(g)| \leq \epsilon/4$ and choose $\delta \leq \frac{h^{m+1}}{C\, s}\frac{\epsilon}{2}$. In the proof of Theorem 3 it is shown that for one function g there exist constants K_1 and K_2 independent of h, s and the function class $\mathcal{F}$ such that the following Bernstein-type inequality holds

$$\mathrm{P}\left(|\tilde{S}_{\lambda,h,n}(g) - \mathbb{E}\, \tilde{S}_{\lambda,h,n}(g)| \geq \frac{\epsilon}{4}\right) \leq 2\, e^{-\frac{[n/2]h^m (1/s)^4 \epsilon^2}{32K_1 + 32/3 \frac{\epsilon}{s^2} K_2}}.$$

Taking the union bound over the covering $C_\delta(\mathcal{F})$ yields

$$\mathrm{P}\left(\sup_{g\in C_\delta(F)} |\tilde{S}_{\lambda,h,n}(g) - \mathbb{E}\, \tilde{S}_{\lambda,h,n}(g)| \geq \frac{\epsilon}{4}\right) \leq 2\mathcal{N}\Big(\delta,\, \mathcal{F},\, \|\cdot\|_\infty\Big) e^{-\frac{[n/2]h^m (1/s)^4 \epsilon^2}{32K_1 + 32/3 \frac{\epsilon}{s^2} K_2}}$$

In total we have on the event $\mathcal{E}_1$ and $\mathcal{E}_2$

$$\sup_{g\in\mathcal{F}} |S_{\lambda,h,n}(g) - \mathbb{E}\,\tilde{S}_{\lambda,h,n}(g)| \leq I + II \leq \frac{\epsilon}{4} + \frac{\epsilon}{2} + \frac{\epsilon}{4} \leq \epsilon$$

Putting the results for $\mathcal{E}_1$ and $\mathcal{E}_2$ together we are done. □

Note that despite $\mathcal{F}$ is not required to be uniformly bounded in the previous theorem, one gets only finite covering numbers in the $\|\cdot\|_\infty$ norm under this condition. In order to get finite sample bounds, we need to know how far $\mathbb{E}\,\tilde{S}_{\lambda,h,n}$ is away from its limit for finite h uniformly over a certain function class $\mathcal{F}$.

Theorem 5. *Let $\mathcal{F}$ be a function class such that $\sup_{f\in\mathcal{F}} \|f\|_{C^3(M)} \leq s$. Then there exist constants $C', C'' > 0$ depending only on M, p and the kernel k such that for all $h < C' \min\{\frac{\pi\rho}{3}, \frac{r_i}{3}, \frac{\kappa}{2R_k}, \frac{R_0}{R_k}\}$,*

$$\sup_{f\in\mathcal{F}} \left| \mathbb{E}\,\tilde{S}_{\lambda,h,n} - \frac{C_2}{2C_1^\lambda} \int_M \langle \nabla f, \nabla f\rangle_{T_xM}\, p(x)^{2-2\lambda} \sqrt{\det g}\, dx \right| \leq C''\, s^2 h$$

Proof: Let us first define

$$B_{\lambda,h}(x) := \frac{1}{2h^2} \int_M (f(x) - f(y))^2 k_h(\|i(x) - i(y)\|) \frac{p(y)}{(p_h(x)p_h(y))^\lambda} dV(y).$$

so that $\mathbb{E}\,\tilde{S}_{\lambda,h,n} = \int_M B_{\lambda,h}(x)\, p(x)\, dV(x)$. Now we decompose M as $M = M\backslash N(r) \cup N(r)$, where $r \leq r_i$ (see Definition 2), which implies that for all $x \in M\backslash N(r)$ there exist normal coordinates on the ball $B_M(x, r)$, that is $\mathrm{inj}(M\backslash N(r)) = r$. The expansion of Proposition 1 holds pointwise for $h_0 \leq \frac{C'}{3}\min\{\pi\rho, \mathrm{inj}(x)\}$[10]. Since ρ is lower-bounded due to compactness of M and $\mathrm{inj}(M\backslash N(r)) = r$ we can use Proposition 1 uniformly over $M\backslash N(r)$, which yields

$$\sup_{x\in M\backslash N(r)} \left| B_{\lambda,h}(x) - \frac{C_2}{2C_1^\lambda} \langle \nabla f, \nabla f\rangle_{T_xM}\, p(x)^{1-2\lambda} \right| \leq C'' s^2 h,$$

where C'' is independent of $\mathcal{F}$. Therefore the bound holds uniformly over the function class $\mathcal{F}$. Next we have two error terms I and II:

$$I := \int_{N(r)} B_{\lambda,h}(x) p(x) dV(x), \quad II := \frac{C_2}{2C_1^\lambda} \int_{N(r)} \|\nabla f\|^2\, p^{2-2\lambda}(x) dV(x)$$

Let us first deal with I. By Lemma 2 we have for $hR_k \leq \frac{\kappa}{2}$, $d_M(x, y) \leq 2\|x - y\| \leq 2hR_k$ (due to compact support of k). Together with the volume bound from Lemma 1 we get for $hR_k \leq \min\{\kappa/2, R_0/2\}$:

$$|B_{\lambda,h}(x)| \leq \frac{2s^2 R_k^2 \|k\|_\infty \|p\|_\infty}{D_1^{2\lambda}} S_2 2^m R_k^m$$

[10] The factor $1/3$ arises since we have to take care that also for all points $y \in B_M(x, r/3)$ we can do the expansion for $p_h(y)$.

Again using the volume bound from Proposition 1 for $r \leq R_0$ yields:

$$I \leq \frac{2s^2 R_k^2 \left\|k\right\|_\infty \left\|p\right\|_\infty^2}{D_1^{2\lambda}} S_2 2^m R_k^m S_2\, r \text{ vol}(\partial M) := C''' s^2\, r$$

By the volume bound and $\left\|\nabla f\right\|_\infty \leq s$ we get $II \leq C'''' s^2 r$. For $r \leq \pi\rho$ we choose $h = C\, r$ for some constant C so that all error terms are of order $s^2\, h$. $\square$

Theorems 4 and 5 together provide a finite sample result for the convergence of $S_{\lambda,h,n}$ over a sufficiently smooth function class $\mathcal{F}$. We use now the upper bounds on the covering numbers of a ball of α-Hölder functions in order to get an explicit finite sample bound and rates for $h(n)$. Moreover we let $s(n) \to \infty$ so that in the limit we get uniform convergence for all α-Hölder functions.

Theorem 6. *Let $\mathcal{F}_\alpha(s)$ be the ball of radius s in the space of Hölder functions $\mathcal{F}_\alpha$ on M. Define $\gamma = 2 - 2\lambda$ and $c = \frac{C_2}{2C_1^\lambda}$, then for $\alpha \geq 3$ and $h \to 0$ and $nh^{\frac{m^2+m+\alpha m}{\alpha}} \to \infty$,*

$$\sup_{f \in \mathcal{F}_\alpha(s)} \left| S_{\lambda,h,n}(f) - c\, S_{\Delta_\gamma}(f) \right| = O\left(\frac{s^2}{(nh^{\frac{m^2+m+\alpha m}{\alpha}})^{\frac{\alpha}{2\alpha+m}}} \right) + O(s^2\, h) \quad \text{a.s.}$$

The optimal rate for h is $h = O(n^{-\frac{\alpha}{2\alpha+2m+m^2+m\alpha}})$.
Let $s = log(n)$, then if $h \to 0$ and $nh^{\frac{m^2+m+\alpha m}{\alpha}}/log(n)^{\frac{4\alpha+2m}{\alpha}} \to \infty$ one has,

$$\forall\, f \in \mathcal{F}_\alpha, \quad \lim_{n\to\infty} S_{\lambda,h,n}(f) = \frac{C_2}{2C_1^\lambda} \int_M \left\|\nabla f\right\|_{T_x M}^2 p(x)^{2-2\lambda} \sqrt{\det g}\, dx, \quad \text{a.s..}$$

Proof: For $\alpha > 3$, we have $\mathcal{F}_\alpha \subset C^3(M)$ and $\left\|f\right\|_{C^3(M)} \leq \left\|f\right\|_{\mathcal{F}_\alpha}, \forall\, f \in \mathcal{F}_\alpha$, so that we can apply Theorems 4 and 5. The first statement follows for sufficiently small h and by plugging the bound on the covering numbers of $\mathcal{F}_\alpha(s)$ from Theorem 2 into Theorem 4 and putting Theorem 4 and 5 together. The dominating terms of $\log \mathrm{P}(\sup_{f\in\mathcal{F}_\alpha(s)} |S_{\lambda,h,n}(f) - \mathbb{E}\, S_{\lambda,h,n}(f)| > \epsilon)$ are

$$\left(\frac{2C\, s^2}{\epsilon\, h^{m+1}}\right)^{\frac{m}{\alpha}} - \frac{nh^m\, (1/s)^4\, \epsilon^2}{4C} = \left(\frac{2C\, s^2}{\epsilon\, h^{m+1}}\right)^{\frac{m}{\alpha}} \left[1 - \frac{nh^{\frac{m^2+m+\alpha m}{\alpha}}\, (1/s)^{4+2\frac{m}{\alpha}}\, \epsilon^{2+\frac{m}{\alpha}}}{C'}\right]$$

so that for the given rate the term in the bracket can be made negative and and the whole term is summable so that almost sure convergence follows by the Borel-Cantelli Lemma. The optimal rate for $h(n)$ can be computed by equating the two order-terms. For the second statement we simply choose $s = log(n)$. $\square$

This theorem provides uniform convergence of the adaptive regularization functional $S_{\lambda,h,n}(f)$ over the large class of α-Hölder functions. We think that this theorem will be helpful to prove consistency results for algorithms which use $S_{\lambda,h,n}(f)$ as a regularizer. As expected the rate depends only on the intrinsic dimension m and not on the extrinsic dimension d. At least for low-dimensional submanifolds we can therefore get a good approximation of the continuous regularization functional even if we work in a high-dimensional space.

References

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comp.*, 15(6):1373–1396, 2003.

[2] M. Belkin and P. Niyogi. Semi-supervised learning on manifolds. *Machine Learning*, 56:209–239, 2004.

[3] M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. In P. Auer and R. Meir, editors, *Proc. of the 18th Conf. on Learning Theory (COLT)*, Berlin, 2005. Springer.

[4] O. Bousquet, O. Chapelle, and M. Hein. Measure based regularization. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Adv. in Neur. Inf. Proc. Syst. (NIPS)*, volume 16. MIT Press, 2004.

[5] S. Canu and A. Elisseeff. Regularization, kernels and sigmoid net. unpublished, 1999.

[6] S. Coifman and S. Lafon. Diffusion maps. Preprint, Jan. 2005, to appear in Appl. and Comp. Harm. Anal., 2005.

[7] E. Hebey. *Nonlinear analysis on manifolds: Sobolev spaces and inequalities*. Courant Institute of Mathematical Sciences, New York, 1998.

[8] M. Hein, J.-Y. Audibert, and U. von Luxburg. From graphs to manifolds - weak and strong pointwise consistency of graph Laplacians. In P. Auer and R. Meir, editors, *Proc. of the 18th Conf. on Learning Theory (COLT)*, Berlin, 2005. Springer.

[9] M. Hein. *Geometrical aspects of statistical learning theory*. PhD thesis, MPI für biologische Kybernetik/Technische Universität Darmstadt, 2005. http://www.kyb.mpg.de/publication.html?user=mh.

[10] H. Hendriks, J.H.M. Janssen, and F.H. Ruymgaart. Strong uniform convergence of density estimators on compact Euclidean manifolds. *Statist. Prob. Lett.*, 16:305–311, 1993.

[11] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[12] A. N. Kolmogorov and V. M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *Amer. Math. Soc. Transl.*, 17:277–364, 1961.

[13] T. Schick. Manifolds with boundary of bounded geometry. *Math. Nachr.*, 223:103–120, 2001.

[14] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer, New-York, second edition, 2001.

[15] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Adv. in Neur. Inf. Proc. Syst. (NIPS)*, volume 16. MIT Press, 2004.

Appendix: Covering numbers for α-Hölder functions on compact Riemannian manifolds (with boundary)

Proof: [Proof of Theorem 1] The first property follows by the maximality of the separated subsets. It remains to prove the upper bounds on the cardinality of I_1 and I_2. The sets $\{n(x'_i, \frac{\epsilon}{2}, \frac{\epsilon}{2})\}_{i\in I_1}$ and $\{n(x_i, \frac{\epsilon}{2})\}_{i\in I_2}$ are disjoint. Therefore

$$\sum_{i\in I_1} \mathrm{vol}\left(n(x'_{i_1}, \frac{\epsilon}{2}, \frac{\epsilon}{2})\right) \leq \mathrm{vol}\left(N(\epsilon)\right), \qquad \sum_{i\in I_2} \mathrm{vol}\left(n(x_{i_2}, \frac{\epsilon}{2})\right) \leq \mathrm{vol}\left(M\right).$$

Then use $\mathrm{vol}\big(N(\epsilon)\big) \leq S_2 \epsilon \,\mathrm{vol}\big(\partial M\big)$ and the volume bounds in Lemma 1. $\square$

Now we are ready to prove the result on covering numbers for $\mathcal{F}_\alpha(s)$.

Proof: [Proof of Theorem 2] Let $\delta = \big(\frac{\epsilon}{3s\, e^{2m}}\big)^{1/\alpha}$ and let $T_1 = \{z_i'\}_{i\in I_1}$ and $T_2 = \{z_i\}_{i\in I_2}$ describe a maximal δ-separated set of ∂M and $M\backslash N(\delta)$ as in Theorem 1. For each vector $k = (k_1, \ldots, k_d)$ with $k \leq \underline{\alpha}$ we form for each $f \in \mathcal{F}_\alpha(s)$ the two vectors

$$A_k f = \left(\Big[\frac{D^k(f\circ\phi_{z_1'})(0)}{s\,\delta^{\alpha-k}}\Big], \ldots, \Big[\frac{D^k(f\circ\phi_{z_{|I_1|}'})(0)}{s\,\delta^{\alpha-k}}\Big] \right)$$
$$B_k f = \left(\Big[\frac{D^k(f\circ\phi_{z_1})(0)}{s\,\delta^{\alpha-k}}\Big], \ldots, \Big[\frac{D^k(f\circ\phi_{z_{|I_2|}})(0)}{s\,\delta^{\alpha-k}}\Big] \right),$$

where $[\cdot]$ denotes rounding to the closest integer and ϕ denotes the normal charts corresponding to the points in T_1 and T_2. Note that the vector $A_k f$ is well-defined since all derivatives of f are uniformly bounded. Now let f_1 and f_2 be two functions such that $A_k f_1 = A_k f_2$ and $B_k f_1 = B_k f_2$ for each $k \leq \underline{\alpha}$. Define $g = f_1 - f_2$, then one has for every $z \in T_1 \cup T_2$

$$|D^k g(z)| = |D^k f_1(z) - D^k f_2(z)| \leq s\,\delta^{\alpha-k} \tag{4}$$

Moreover for every $x \in M\backslash N(\delta)$ there exists an $z_i \in T_2$ such that $d(x, z_i) \leq \delta$ and for every $x \in N(\delta)$ there exists an $z_i' \in T_1$ such that $d(x, z_i') \leq 2\delta$ (this follows from the definition of normal collar charts and the triangle inequality[11]). Since $M \subset M\backslash N(\delta) \cup N(\delta)$ there exists for each $x \in M$ a corresponding normal chart ϕ_z based on $z \in N_1 \cup N_2$ such that for each coordinate $x_i = (\phi_z^{-1}(x))_i$, $i = 1, \ldots, m$ of x one has $x_i \leq \max\{\delta, 2\delta\} = 2\delta$. Now we do a Taylor expansion of g around $z = \phi_z(0)$ in the normal chart ϕ_z and get for $x = \phi_z((x_1, \ldots, x_m))$:

$$g(x) = \sum_{k\leq\underline{\alpha}} D^k(g\circ\phi_z)(0) \prod_{i=1}^m \frac{x_i^{k_i}}{k_i!} + \sum_{k=\underline{\alpha}} \Big(D^k(g\circ\phi_z)(\lambda x_i) - D^k(g\circ\phi_z)(0) \Big) \prod_{i=1}^m \frac{x_i^{k_i}}{k_i!}$$

with $\lambda \in [0,1]$. By (4) and the Lipschitz property of functions in $\mathcal{F}_\alpha(s)$ we get

$$|g(x)| \leq \sum_{k\leq\underline{\alpha}} s\delta^{\alpha-k}\frac{(2\delta)^k}{k!} + 2s\frac{m^{\underline{\alpha}}}{\underline{\alpha}!}2^\alpha\delta^\alpha \leq \delta^\alpha e^{2m}(s+2s) \leq 3se^{2m}\delta^\alpha = \epsilon,$$

so that the covering numbers of an ϵ-covering of $\mathcal{F}_\alpha(s)$ are upper bounded by the number of possible matrices Af and Bf for $f \in \mathcal{F}_\alpha(s)$. The number of possible derivatives $\leq \underline{\alpha}$ is upper bounded by $\sum_{i=0}^{\underline{\alpha}} m^i = \frac{m^{\underline{\alpha}}-1}{m-1}$ for $m > 1$ and α for $m = 1$. Since in $\mathcal{F}_\alpha(s)$ the derivatives fulfill $|D^k f(x)| \leq s$ for each k, $A_k f$ contains $\frac{2}{\delta^{\alpha-k}} + 2$ values which is upper bounded by $\frac{2}{\delta^\alpha} + 2$. Thus for one point in the covering the number of different values in Af is upper bounded by

[11] One follows the geodesic along the boundary which is shorter than δ and then the geodesic along the inward normal vector which has also length shorter than δ.

$\left(\frac{2}{\delta^\alpha}+2\right)^{2m^\alpha}$ for $m \geq 2$ and $\left(\frac{2}{\delta^\alpha}+2\right)^{\alpha}$ for $m = 1$. The same holds for $B_k f$. Assume now we reorder the set N_1 in such a way that for each $j > 1$ there is an index $i < j$ such that $d(z'_i, z'_j) \leq 2\delta$. We compute now the range over which values of $A_k f(z'_j)$ vary given the values of $A_k f(z'_i)$. The problem is that the derivatives of f at z'_j and z'_i are given with respect to different normal charts $\phi_{z'_j}$ and $\phi_{z'_i}$. In order to compare $A_k f(z'_j)$ with $A_k f(z'_i)$ we therefore have to change coordinates. Let x^μ be the coordinates with respect to $\phi_{z'_j}$ and y^μ with respect to $\phi_{z'_i}$. Then one has e.g. for the second derivative,

$$\frac{\partial^2 f}{\partial x_\mu \partial x_\nu} = \frac{\partial^2 f}{\partial y_\beta \partial y_\gamma}\frac{\partial y_\beta}{\partial x_\mu}\frac{\partial y_\gamma}{\partial x_\nu} + \frac{\partial f}{\partial y_\alpha}\frac{\partial^2 y_\alpha}{\partial x_\mu \partial x_\nu} =: C_y^2 f,$$

with the obvious generalization to higher orders. By Taylor's theorem one gets

$$D_x^k f(x_j) = D_x^k (f \circ \phi_{z'_j})(0) = \sum_{k+l \leq \underline{\alpha}} D_x^{k+l}(f \circ \phi_{z'_j})(x^i)\frac{x^l}{l!} + R = \sum_{k+l \leq \underline{\alpha}} C_y^{k+l}(0)\frac{x^l}{l!} + R$$

Define $B_y^{k+l}(0)$ as $C_y^{k+l}(0)$ with derivatives replaced by their discretized values,

$$\frac{\partial^k f}{\partial y_{i_1} \ldots \partial y_{i_k}}(0) \longrightarrow s\,\delta^{\alpha-k}\left[\frac{\partial^k f}{\partial y_{i_1} \ldots \partial y_{i_k}}(0)\frac{1}{s\,\delta^{\alpha-k}}\right]$$

Given now all the discretized values $Af(z'_i)$ we arrive at

$$\left| D_x^k (f \circ \phi_{z'_j})(0) - \sum_{k+l \leq \underline{\alpha}} B_y^{k+l}(0)\frac{x^l}{l!}\right| \leq \sum_{k+l \leq \underline{\alpha}} \left| C_y^{k+l}(0) - B_y^{k+l}(0)\right|\frac{x^l}{l!} + |R|$$

The leading term of the summands can be upper bounded as follows

$$\left| C_y^{k+l}(0) - B_y^{k+l}(0)\right| \leq s\,(\Gamma\, m)^k \delta^{\alpha-k-l}$$

where $\Gamma = \max_{i,j} \max_{k \leq \underline{\alpha}} \sup_{x \in M} D^k(\phi_{z'_i}^{-1} \circ \phi_{z'_j})$. It can be shown that the remainder term $|R|$ is of order $s\,\delta^{\alpha-k}$, so that in total we get that there exists a constant C depending on Γ, m and α such that

$$\left| D_x^k (f \circ \phi_{z'_j})(0) - \sum_{k+l \leq \underline{\alpha}} B_y^{k+l}(0)\frac{x^l}{l!}\right| \leq C\, s\, \delta^{\alpha-k}$$

That implies that given the values of Af at x_i the values of Af at x_j vary over an interval of size $C\, s\, \frac{\delta^{\alpha-k}}{\delta^{\alpha-k}} = C\, s$. Using our previous bound on the number of possible values of Af for one point we get that the total number of values of Af is upper bounded as follows:

$$|Af| \leq \left(\frac{2}{\delta^\alpha}+2\right)^{2m^\alpha} \left((C\, s)^{2m^\alpha}\right)^{|I_1|}$$

The same can be done for $|Bf|$. Replacing $|I_1|$ resp. $|I_2|$ with the numbers from Theorem 1 and upper bounding $\log(1/\epsilon)$ by $(1/\epsilon)^{m/\alpha}$ finishes the proof. $\square$