Chapter 4

Learning Principal Curves with a Length Constraint

An unfortunate property of the HS definition is that in general, it is not known if principal curves exist for a given source density. This also makes it difficult to theoretically analyze any estimation scheme for principal curves. In Section 4.1 we propose a new concept of principal curves and prove their existence in the new sense for a large class of source densities. In Section 4.2 we consider the problem of principal curve design based on training data. We introduce and analyze an estimation scheme using a common model in statistical learning theory.

4.1 Principal Curves with a Length Constraint

One of the defining properties of the first principal component line is that it minimizes the distance function (18) among all straight lines (Property 2 in Section 2.2.3). We wish to generalize this property of the first principal component and define principal curves so that they minimize the expected squared distance over a class of curves rather than only being critical points of the distance function. To do this it is necessary to constrain the length of the curve since otherwise for any $\mathbf{X}$ with a density and any $\varepsilon > 0$ there exists a smooth curve $\mathbf{f}$ such that $\Delta(\mathbf{f}) \leq \varepsilon$, and thus a minimizing $\mathbf{f}$ has infinite length. On the other hand, if the distribution of $\mathbf{X}$ is concentrated on a polygonal line and is uniform there, the infimum of the squared distances $\Delta(\mathbf{f})$ is 0 over the class of smooth curves but no smooth curve can achieve this infimum. For this reason, we relax the requirement that $\mathbf{f}$ should be differentiable but instead we constrain the length of $\mathbf{f}$. Note that by the definition of curves, $\mathbf{f}$ is still continuous. We give the following new definition of principal curves.

Definition 4 *A curve* $\mathbf{f}^*$ *is called a* principal curve *of length* L *for* $\mathbf{X}$ *if* $\mathbf{f}^*$ *minimizes* $\Delta(\mathbf{f})$ *over all curves of length less than or equal to* L*.*

The relation of our definition and the HS definition (Definition2) is analogous to the relation of a globally optimal vector quantizer and a locally optimal vector quantizer (Section 2.1). Locally optimal vector quantizers are fixed points of the expected distortion $\Delta(q)$ while self-consistent principal curves are fixed points of the distance function $\Delta(\mathbf{f})$. This similarity is further illuminated by a recent work [TLF95] which defines k points $\mathbf{y}_1,\ldots,\mathbf{y}_k$ to be self-consistent if

$$\mathbf{y}_i = E[\mathbf{X}|\mathbf{X}\in V_i]$$

where $V_1,\ldots,V_k$ are the Voronoi regions associated with $\mathbf{y}_1,\ldots,\mathbf{y}_k$. In this sense, our principal curves correspond to globally optimal vector quantizers ("principal points" by the terminology of [TLF95]) while the HS principal curves correspond to self-consistent points.

A useful advantage of the new definition is that principal curves of length L always exist if $\mathbf{X}$ has finite second moments as the next result shows.

Theorem 1 *Assume that $E\|\mathbf{X}\|^2<\infty$. Then for any $L>0$ there exists a curve $\mathbf{f}^*$ with $l(\mathbf{f}^*)\le L$ such that*

$$\Delta(\mathbf{f}^*)=\inf\{\Delta(\mathbf{f}):l(\mathbf{f})\le L\}.$$

Proof Define

$$\Delta^*=\inf\{\Delta(\mathbf{f}):l(\mathbf{f})\le L\}.$$

First we show that the above infimum does not change if we add the restriction that all $\mathbf{f}$ lie inside a closed sphere $S(r)=\{\mathbf{x}:\|\mathbf{x}\|\le r\}$ of large enough radius r and centered at the origin. Indeed, without excluding nontrivial cases, we can assume that $\Delta^*<E\|\mathbf{X}\|^2$. Denote the distribution of $\mathbf{X}$ by μ and choose $r>3L$ large enough such that

$$\int_{S(r/3)}\|\mathbf{x}\|^2\mu(d\mathbf{x})>\Delta^*+\varepsilon \tag{46}$$

for some $\varepsilon>0$. If $\mathbf{f}$ is such that $G_{\mathbf{f}}$ (the graph of $\mathbf{f}$ defined by 17) is not entirely contained in $S(r)$, then for all $\mathbf{x}\in S(r/3)$ we have $\Delta(\mathbf{x},\mathbf{f})>\|\mathbf{x}\|^2$ since the diameter of $G_{\mathbf{f}}$ is at most L. Then (46) implies that

$$\Delta(\mathbf{f})\ge\int_{S(r/3)}\Delta(\mathbf{x},\mathbf{f})\mu(d\mathbf{x})>\Delta^*+\varepsilon$$

and thus

$$\Delta^*=\inf\{\Delta(\mathbf{f}):l(\mathbf{f})\le L,G_{\mathbf{f}}\subset S(r)\}. \tag{47}$$

In view of (47) there exists a sequence of curves $\{\mathbf{f}_n\}$ such that $l(\mathbf{f}_n)\le L$, $G_{\mathbf{f}_n}\subset S(r)$ for all n, and $\Delta(\mathbf{f}_n)\to\Delta^*$. By the discussion preceding (16) in Section 2.2.1, we can assume without loss of generality that all $\mathbf{f}_n$ are defined over $[0,1]$ and

$$\|\mathbf{f}_n(t_1)-\mathbf{f}_n(t_2)\|\le L|t_1-t_2| \tag{48}$$

for all $t_1, t_2 \in [0,1]$. Consider the set of all curves $\mathcal{C}$ over $[0,1]$ such that $\mathbf{f} \in \mathcal{C}$ if and only if $\|\mathbf{f}(t_1) - \mathbf{f}(t_2)\| \leq L|t_1 - t_2|$ for all $t_1, t_2 \in [0,1]$ and $G_{\mathbf{f}} \subset S(r)$. It is easy to see that $\mathcal{C}$ is a closed set under the uniform metric $d(\mathbf{f}, \mathbf{g}) = \sup_{0 \leq t \leq 1} \|\mathbf{f}(t) - \mathbf{g}(t)\|$. Also, $\mathcal{C}$ is an equicontinuous family of functions and $\sup_t \|\mathbf{f}(t)\|$ is uniformly bounded over $\mathcal{C}$. Thus $\mathcal{C}$ is a compact metric space by the Arzela-Ascoli theorem (see, e.g., [Ash72]). Since $\mathbf{f}_n \in \mathcal{C}$ for all n, it follows that there exists a subsequence $\mathbf{f}_{n_k}$ converging uniformly to an $\mathbf{f}^* \in \mathcal{C}$.

To simplify the notation let us rename $\{\mathbf{f}_{n_k}\}$ as $\{\mathbf{f}_n\}$. Fix $\mathbf{x} \in \mathbb{R}^d$, assume $\Delta(\mathbf{x}, \mathbf{f}_n) \geq \Delta(\mathbf{x}, \mathbf{f}^*)$, and let $t_{\mathbf{x}}$ be such that $\Delta(\mathbf{x}, \mathbf{f}^*) = \|\mathbf{x} - \mathbf{f}^*(t_{\mathbf{x}})\|^2$. Then by the triangle inequality,

$$\begin{aligned} |\Delta(\mathbf{x}, \mathbf{f}^*) - \Delta(\mathbf{x}, \mathbf{f}_n)| &= \Delta(\mathbf{x}, \mathbf{f}_n) - \Delta(\mathbf{x}, \mathbf{f}^*) \\ &\leq \|\mathbf{x} - \mathbf{f}_n(t_{\mathbf{x}})\|^2 - \|\mathbf{x} - \mathbf{f}^*(t_{\mathbf{x}})\|^2 \\ &\leq (\|\mathbf{x} - \mathbf{f}_n(t_{\mathbf{x}})\| + \|\mathbf{x} - \mathbf{f}^*(t_{\mathbf{x}})\|) \, \|\mathbf{f}_n(t_{\mathbf{x}}) - \mathbf{f}^*(t_{\mathbf{x}})\|. \end{aligned}$$

By symmetry, a similar inequality holds if $\Delta(\mathbf{x}, \mathbf{f}_n) < \Delta(\mathbf{x}, \mathbf{f}^*)$. Since $G_{\mathbf{f}^*}, G_{\mathbf{f}_n} \subset S(r)$, and $E\|\mathbf{X}\|^2$ is finite, there exists $A > 0$ such that

$$E\,|\Delta(\mathbf{X}, \mathbf{f}_n) - \Delta(\mathbf{X}, \mathbf{f}^*)| \leq A \sup_{0 \leq t \leq 0} \|\mathbf{f}_n(t) - \mathbf{f}^*(t)\|$$

and therefore

$$\Delta^* = \lim_{n \to \infty} \Delta(\mathbf{f}_n) = \Delta(\mathbf{f}^*).$$

Since the Lipschitz condition on $\mathbf{f}^*$ guarantees that $l(\mathbf{f}^*) \leq L$, the proof is complete. □

Note that we have dropped the requirement of the HS definition that principal curves be non-intersecting. In fact, Theorem 1 does not hold in general for non-intersecting curves of length L without further restricting the distribution of $\mathbf{X}$ since there are distributions for which the minimum of $\Delta(\mathbf{f})$ is achieved only by an intersecting curve even though non-intersecting curves can arbitrarily approach this minimum. Note also that neither the HS nor our definition guarantees the uniqueness of principal curves. In our case, there might exist several principal curves for a given length constraint L but each of these will have the same (minimal) squared loss.

Finally, we note that although principal curves of a given length always exist, it appears difficult to demonstrate concrete examples unless the distribution of $\mathbf{X}$ is discrete or it is concentrated on a curve. It is presently unknown what principal curves look like with a length constraint for even the simplest continuous multivariate distributions such as the Gaussian. However, this fact in itself does not limit the operational significance of principal curves. The same problem occurs in the theory of optimal vector quantizers (Section 2.1.1) where, except for the scalar case ($d = 1$), the structure of optimal quantizers with $k > 2$ codepoints is unknown for even the most common multivariate densities. Nevertheless, algorithms for quantizer design attempting to find near optimal vector quantizers are of great theoretical and practical interest.

4.2 Learning Principal Curves

Suppose that n independent copies $\mathbf{X}_1, \ldots, \mathbf{X}_n$ of $\mathbf{X}$ are given. These are called the *training data* and they are assumed to be independent of $\mathbf{X}$. The goal is to use the training data to construct a curve of length at most L whose expected squared loss is close to that of a principal curve for $\mathbf{X}$.

Our method is based on a common model in statistical learning theory (e.g., see [Vap98]). We consider classes $\S_1, \S_2, \ldots$ of curves of increasing complexity. Given n data points drawn independently from the distribution of $\mathbf{X}$, we choose a curve as the estimator of the principal curve from the kth model class $\S_k$ by minimizing the empirical error. By choosing the complexity of the model class appropriately as the size of the training data grows, the chosen curve represents the principal curve with increasing accuracy.

We assume that the distribution of $\mathbf{X}$ is concentrated on a closed and bounded convex set $K \subset \mathbb{R}^d$. The following lemma shows that there exists a principal curve of length L *inside* K, and so we will only consider curves in K.

Lemma 1 *Assume that $P\{X \in K\} = 1$ for a closed and convex set K, and let $\mathbf{f}$ be a curve with $l(\mathbf{f}) \leq L$. Then there exists a curve $\hat{\mathbf{f}}$ such that $G_{\hat{\mathbf{f}}} \subset K$, $l(\hat{\mathbf{f}}) \leq L$, and*

$$\Delta(\hat{\mathbf{f}}) \leq \Delta(\mathbf{f}).$$

Proof For each t in the domain of $\mathbf{f}$, let $\hat{\mathbf{f}}(t)$ be the unique point in K such that $\|\mathbf{f}(t) - \hat{\mathbf{f}}(t)\| = \min_{\mathbf{x} \in K} \|\mathbf{f}(t) - \mathbf{x}\|$. It is well known that $\hat{\mathbf{f}}(t)$ satisfies

$$(\mathbf{f}(t) - \hat{\mathbf{f}}(t))^T (\mathbf{x} - \hat{\mathbf{f}}(t)) \leq 0, \quad \text{for all } \mathbf{x} \in K. \tag{49}$$

Then for all t_1, t_2 we have

$$\begin{aligned} \|\mathbf{f}(t_1) - \mathbf{f}(t_2)\|^2 &= \|\hat{\mathbf{f}}(t_1) - \hat{\mathbf{f}}(t_2)\|^2 + \|\mathbf{f}(t_1) - \hat{\mathbf{f}}(t_1) + \hat{\mathbf{f}}(t_2) - \mathbf{f}(t_2)\|^2 + \\ &\quad 2(\hat{\mathbf{f}}(t_1) - \hat{\mathbf{f}}(t_2))^T (\mathbf{f}(t_1) - \hat{\mathbf{f}}(t_1)) + 2(\hat{\mathbf{f}}(t_1) - \hat{\mathbf{f}}(t_2))^T (\hat{\mathbf{f}}(t_2) - \mathbf{f}(t_2)) \\ &\geq \|\hat{\mathbf{f}}(t_1) - \hat{\mathbf{f}}(t_2)\|^2 \end{aligned}$$

where the inequality follows from (49) since $\hat{\mathbf{f}}(t_1), \hat{\mathbf{f}}(t_2) \in K$. Thus $\hat{\mathbf{f}}(t)$ is continuous (it is a curve) and $l(\hat{\mathbf{f}}) \leq l(\mathbf{f}) \leq L$. A similar inequality shows that for all t and $\mathbf{x} \in K$,

$$\|\mathbf{x} - \hat{\mathbf{f}}(t)\|^2 \leq \|\mathbf{x} - \mathbf{f}(t)\|^2$$

so that $\Delta(\hat{\mathbf{f}}) \leq \Delta(\mathbf{f})$. □

Let $\S$ denote the family of curves taking values in K and having length not greater than L. For $k \geq 1$ let $\S_k$ be the set of polygonal (piecewise linear) curves in K which have k segments and whose lengths do not exceed L. Note that $\S_k \subset \S$ for all k. Let $\Delta(\mathbf{x}, \mathbf{f})$ denote the squared distance between

a point $\mathbf{x} \in \mathbb{R}^d$ and the curve $\mathbf{f}$ as defined in (15). For any $\mathbf{f} \in \S$ the empirical squared error of $\mathbf{f}$ on the training data is the sample average

$$\Delta_n(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n} \Delta(\mathbf{X}_i, \mathbf{f}) \tag{50}$$

where we have suppressed in the notation the dependence of $\Delta_n(\mathbf{f})$ on the training data. Let our theoretical algorithm[1] choose an $\mathbf{f}_{k,n} \in \S_k$ which minimizes the empirical error, i.e,

$$\mathbf{f}_{k,n} = \arg\min_{\mathbf{f}\in\S_k} \Delta_n(\mathbf{f}). \tag{51}$$

We measure the efficiency of $\mathbf{f}_{k,n}$ in estimating $\mathbf{f}^*$ by the difference $J(\mathbf{f}_{k,n})$ between the expected squared loss of $\mathbf{f}_{k,n}$ and the optimal expected squared loss achieved by $\mathbf{f}^*$, i.e., we let

$$J(\mathbf{f}_{k,n}) = \Delta(\mathbf{f}_{k,n}) - \Delta(\mathbf{f}^*) = \Delta(\mathbf{f}_{k,n}) - \min_{\mathbf{f}\in\S} \Delta(\mathbf{f}).$$

Since $\S_k \subset \S$, we have $J(\mathbf{f}_{k,n}) \geq 0$. Our main result in this chapter proves that if the number of data points n tends to infinity, and k is chosen to be proportional to $n^{1/3}$, then $J(\mathbf{f}_{k,n})$ tends to zero at a rate $J(\mathbf{f}_{k,n}) = O(n^{-1/3})$.

Theorem 2 *Assume that $P\{\mathbf{X} \in K\} = 1$ for a bounded and closed convex set K, let n be the number of training points, and let k be chosen to be proportional to $n^{1/3}$. Then the expected squared loss of the empirically optimal polygonal line with k segments and length at most L converges, as $n \to \infty$, to the squared loss of the principal curve of length L at a rate*

$$J(\mathbf{f}_{k,n}) = O(n^{-1/3}).$$

The proof of the theorem is given below. To establish the result we use techniques from statistical learning theory (e.g., see [DGL96]). First, the approximating capability of the class of curves $\S_k$ is considered, and then the estimation (generalization) error is bounded via covering the class of curves $\S_k$ with ε accuracy (in the squared distance sense) by a discrete set of curves. When these two bounds are combined, one obtains

$$J(\mathbf{f}_{k,n}) \leq \sqrt{\frac{kC(L,D,d)}{n}} + \frac{DL+2}{k} + O(n^{-1/2}) \tag{52}$$

where the term $C(L,D,d)$ depends only on the dimension d, the length L, and the diameter D of the support of $\mathbf{X}$, but is independent of k and n. The two error terms are balanced by choosing k to be proportional to $n^{1/3}$ which gives the convergence rate of Theorem 2.

[1] The term "hypothetical algorithm" might appear to be more accurate since we have not shown that an algorithm for finding $\mathbf{f}_{k,n}$ exists. However, an algorithm clearly exists which can approximate $\mathbf{f}_{k,n}$ with arbitrary accuracy in a finite number of steps (consider polygonal lines whose vertices are restricted to a fine rectangular grid). The proof of Theorem 2 shows that such approximating curves can replace $\mathbf{f}_{k,n}$ in the analysis.

Remarks

1. Although the constant hidden in the O notation depends on the dimension d, the exponent of n is dimension-free. This is not surprising in view of the fact that the class of curves § is equivalent in a certain sense to the class of Lipschitz functions $\mathbf{f}:[0,1]\rightarrow K$ such that $\|\mathbf{f}(x)-\mathbf{f}(y)\|\leq L|x-y|$ (see (16) in Section 2.2.1). It is known that the ε-entropy, defined by the logarithm of the ε covering number, is roughly proportional to $1/\varepsilon$ for such function classes [KT61]. Using this result, the convergence rate $O(n^{-1/3})$ can be obtained by considering ε-covers of § directly (without using the model classes $\S_k$) and picking the empirically optimal curve in this cover. The use of the classes $\S_k$ has the advantage that they are directly related to the practical implementation of the algorithm given in the next section.

2. Even though Theorem 2 is valid for any given length constraint L, the theoretical algorithm itself gives little guidance about how to choose L. This choice depends on the particular application and heuristic considerations are likely to enter here. One example is given in Chapter 5 where a practical implementation of the polygonal line algorithm is used to recover a "generating curve" from noisy observations.

3. The proof of Theorem 2 also provides information on the distribution of the expected squared error of $\mathbf{f}_{k,n}$ given the training data $\mathbf{X}_1,\ldots,\mathbf{X}_n$. In particular, it is shown at the end of the proof that for all n and k, and δ such that $0<\delta<1$, with probability at least $1-\delta$ we have

$$E\left[\Delta(\mathbf{X},\mathbf{f}_{k,n})|\mathbf{X}_1,\ldots,\mathbf{X}_n\right]-\Delta(\mathbf{f}^*)\leq\sqrt{\frac{kC(L,D,d)-D^4\log(\Delta/2)}{n}}+\frac{DL+2}{k} \qquad (53)$$

   where log denotes natural logarithm and $C(L,D,d)$ is the same constant as in (52).

4. Recently, Smola et al. [SWS98] obtained $O(n^{-1/(2+\alpha)})$ convergence rate using a similar but more general model where the value of α depends on the particular regularizer used in the model. [SWS98] pointed out that although there exist regularizers with $\alpha<1$, in the particular case of a length constraint, $\alpha=2$ so the obtained convergence rate is $O(n^{-1/4})$.

Proof of Theorem 2 Let $\mathbf{f}_k^*$ denote the curve in $\S_k$ minimizing the squared loss, i.e.,

$$\mathbf{f}_k^*=\arg\min_{\mathbf{f}\in\S_k}\Delta(\mathbf{f}).$$

The existence of a minimizing $\mathbf{f}_k^*$ can easily be shown using a simpler version of the proof of Lemma 1. Then $J(\mathbf{f}_{k,n})$ can be decomposed as

$$J(\mathbf{f}_{k,n})=(\Delta(\mathbf{f}_{k,n})-\Delta(\mathbf{f}_k^*))+(\Delta(\mathbf{f}_k^*)-\Delta(\mathbf{f}^*))$$

where, using standard terminology, $\Delta(\mathbf{f}_{k,n}) - \Delta(\mathbf{f}_k^*)$ is called the *estimation error* and $\Delta(\mathbf{f}_k^*) - \Delta(\mathbf{f}^*)$ is called the *approximation error*. We consider these terms separately first, and then choose k as a function of the training data size n to balance the obtained upper bounds in an asymptotically optimal way.

Approximation Error

For any two curves $\mathbf{f}$ and $\mathbf{g}$ of finite length define their (nonsymmetric) distance by

$$\rho(\mathbf{f},\mathbf{g}) = \max_t \min_s \|\mathbf{f}(t) - \mathbf{g}(s)\|.$$

Note that $\rho(\hat{\mathbf{f}},\hat{\mathbf{g}}) = \rho(\mathbf{f},\mathbf{g})$ if $\hat{\mathbf{f}} \sim \mathbf{f}$ and $\hat{\mathbf{g}} \sim \mathbf{g}$, i.e., $\rho(\mathbf{f},\mathbf{g})$ is independent of the particular choice of the parameterization within equivalence classes. Next we observe that if the diameter of K is D, and $G_{\mathbf{f}}, G_{\mathbf{g}} \in K$, then for all $\mathbf{x} \in K$,

$$\Delta(\mathbf{x},\mathbf{g}) - \Delta(\mathbf{x},\mathbf{f}) \leq 2D\rho(\mathbf{f},\mathbf{g}), \tag{54}$$

and therefore

$$\Delta(\mathbf{g}) - \Delta(\mathbf{f}) \leq 2D\rho(\mathbf{f},\mathbf{g}). \tag{55}$$

To prove (54), let $\mathbf{x} \in K$ and choose t' and s' such that $\Delta(\mathbf{x},\mathbf{f}) = \|\mathbf{x} - \mathbf{f}(t')\|^2$ and $\min_s \|\mathbf{g}(s) - \mathbf{f}(t')\| = \|\mathbf{g}(s') - \mathbf{f}(t')\|$. Then

$$\begin{aligned}
\Delta(\mathbf{x},\mathbf{g}) - \Delta(\mathbf{x},\mathbf{f}) &\leq \|\mathbf{x} - \mathbf{g}(s')\|^2 - \|\mathbf{x} - \mathbf{f}(t')\|^2 \\
&= \left(\|\mathbf{x} - \mathbf{g}(s')\| + \|\mathbf{x} - \mathbf{f}(t')\|\right)\left(\|\mathbf{x} - \mathbf{g}(s')\| - \|\mathbf{x} - \mathbf{f}(t')\|\right) \\
&\leq 2D\|\mathbf{g}(s') - \mathbf{f}(t')\| \\
&\leq 2D\rho(\mathbf{f},\mathbf{g}).
\end{aligned}$$

Let $\mathbf{f} \in \S$ be an arbitrary arc length parameterized curve over $[0, L']$ where $L' \leq L$. Define $\mathbf{g}$ as a polygonal curve with vertices $\mathbf{f}(0), \mathbf{f}(L'/k), \ldots, \mathbf{f}((k-1)L'/k), \mathbf{f}(L')$. For any $t \in [0, L']$, we have $|t - iL'/k| \leq L/(2k)$ for some $i \in \{0, \ldots, k\}$. Since $\mathbf{g}(s) = \mathbf{f}(iL'/k)$ for some s, we have

$$\begin{aligned}
\min_s \|\mathbf{f}(t) - \mathbf{g}(s)\| &\leq \|\mathbf{f}(t) - \mathbf{f}(iL'/k)\| \\
&\leq |t - iL'/k| \leq \frac{L}{2k}.
\end{aligned}$$

Note that $l(\mathbf{g}) \leq L'$, by construction, and thus $\mathbf{g} \in \S_k$. Thus for every $\mathbf{f} \in \S$ there exists a $\mathbf{g} \in \S_k$ such that $\rho(\mathbf{f},\mathbf{g}) \leq L/(2k)$. Now let $\mathbf{g} \in \S_k$ be such that $\rho(\mathbf{f}^*,\mathbf{g}) \leq L/(2k)$. Then by (55) we conclude that the approximation error is upper bounded as

$$\begin{aligned}
\Delta(\mathbf{f}_k^*) - \Delta(\mathbf{f}^*) &\leq \Delta(\mathbf{g}) - \Delta(\mathbf{f}^*) \\
&\leq 2D\rho(\mathbf{f}^*,\mathbf{g}) \\
&\leq \frac{DL}{k}.
\end{aligned} \tag{56}$$

Estimation Error

For each $\varepsilon > 0$ and $k \geq 1$ let $S_{k,\varepsilon}$ be a *finite* set of curves in K which form an ε-cover of $\S_k$ in the following sense. For any $\mathbf{f} \in \S_k$ there is an $\mathbf{f}' \in \S_{k,\varepsilon}$ which satisfies

$$\sup_{\mathbf{x} \in K} |\Delta(\mathbf{x}, \mathbf{f}) - \Delta(\mathbf{x}, \mathbf{f}')| \leq \varepsilon. \tag{57}$$

The explicit construction of $S_{k,\varepsilon}$ is given below in Lemma 2. Since $\mathbf{f}_{k,n} \in \S_k$ (see (51)), there exists an $\mathbf{f}'_{k,n} \in \S_{k,\varepsilon}$ such that $|\Delta(\mathbf{x}, \mathbf{f}_{k,n}) - \Delta(\mathbf{x}, \mathbf{f}'_{k,n})| \leq \varepsilon$ for all $\mathbf{x} \in K$. We introduce the compact notation $\mathcal{X}_n = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ for the training data. Thus we can write

$$\begin{aligned}
E[\Delta(\mathbf{X}, \mathbf{f}_{k,n}) | \mathcal{X}_n] - \Delta(\mathbf{f}_k^*) &= E[\Delta(\mathbf{X}, \mathbf{f}_{k,n}) | \mathcal{X}_n] - \Delta_n(\mathbf{f}_{k,n}) + \Delta_n(\mathbf{f}_{k,n}) - \Delta(\mathbf{f}_k^*) \\
&\leq 2\varepsilon + E[\Delta(\mathbf{X}, \mathbf{f}'_{k,n}) | \mathcal{X}_n] - \Delta_n(\mathbf{f}'_{k,n}) + \Delta_n(\mathbf{f}_{k,n}) - \Delta(\mathbf{f}_k^*) \qquad (58) \\
&\leq 2\varepsilon + E[\Delta(\mathbf{X}, \mathbf{f}'_{k,n}) | \mathcal{X}_n] - \Delta_n(\mathbf{f}'_{k,n}) + \Delta_n(\mathbf{f}_k^*) - \Delta(\mathbf{f}_k^*) \qquad (59) \\
&\leq 2\varepsilon + 2 \cdot \max_{\mathbf{f} \in \S_{k,\varepsilon} \cup \{\mathbf{f}^*\}} |\Delta(\mathbf{f}) - \Delta_n(\mathbf{f})| \qquad (60)
\end{aligned}$$

where (58) follows from the approximating property of $\mathbf{f}'_{k,n}$ and the fact that the distribution of $\mathbf{X}$ is concentrated on K. (59) holds because $\mathbf{f}_{k,n}$ minimizes $\Delta_n(\mathbf{f})$ over all $\mathbf{f} \in \S_k$, and (60) follows because given $\mathcal{X}_n = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$, $E[\Delta(\mathbf{X}, \mathbf{f}'_{k,n}) | \mathcal{X}_n]$ is an ordinary expectation of the type $E[\Delta(\mathbf{X}, \mathbf{f})]$, $\mathbf{f} \in \S_{k,\varepsilon}$. Thus, for any $t > 2\varepsilon$ the union bound implies

$$\begin{aligned}
&P\{E[\Delta(\mathbf{X}, \mathbf{f}_{k,n}) | \mathcal{X}_n] - \Delta(\mathbf{f}_k^*) > t\} \\
&\qquad \leq P\left\{ \max_{\mathbf{f} \in \S_{k,\varepsilon} \cup \{\mathbf{f}^*\}} |\Delta(\mathbf{f}) - \Delta_n(\mathbf{f})| > \frac{t}{2} - \varepsilon \right\} \\
&\qquad \leq (|S_{k,\varepsilon}| + 1) \max_{\mathbf{f} \in \S_{k,\varepsilon} \cup \{\mathbf{f}^*\}} P\left\{ |\Delta(\mathbf{f}) - \Delta_n(\mathbf{f})| > \frac{t}{2} - \varepsilon \right\} \qquad (61)
\end{aligned}$$

where $|\S_{k,\varepsilon}|$ denotes the cardinality of $\S_{k,\varepsilon}$.

Recall now Hoeffding's inequality [Hoe63] which states that if $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed real random variables such that $0 \leq Y_i \leq A$ with probability one, then for all $u > 0$,

$$P\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} Y_i - E[Y_1] \right| > u \right\} \leq 2e^{-2nu^2/A^2}.$$

Since the diameter of K is D, we have $\|\mathbf{x} - \mathbf{f}(t)\|^2 \leq D^2$ for all $\mathbf{x} \in K$ and $\mathbf{f}$ such that $G_{\mathbf{f}} \in K$. Thus $0 \leq \Delta(\mathbf{X}, \mathbf{f}) \leq D^2$ with probability one and by Hoeffding's inequality, for all $\mathbf{f} \in \S_{k,\varepsilon} \cup \{\mathbf{f}^*\}$ we have

$$P\left\{ |\Delta(\mathbf{f}) - \Delta_n(\mathbf{f})| > \frac{t}{2} - \varepsilon \right\} \leq 2e^{-2n((t/2)-\varepsilon)^2/D^4}$$

which implies by (61) that

$$P\{E[\Delta(\mathbf{X}, \mathbf{f}_{k,n}) | \mathcal{X}_n] - \Delta(\mathbf{f}_k^*) > t\} \leq 2\,(|S_{k,\varepsilon}| + 1)\, e^{-2n((t/2)-\varepsilon)^2/D^4} \tag{62}$$

for any $t > 2\varepsilon$. Using the fact that $E[Y] = \int_0^\infty P\{Y > t\}\,dt$ for any nonnegative random variable Y, we can write for any $u > 0$,

$$\begin{aligned}
\Delta(\mathbf{f}_{k,n}) - \Delta(\mathbf{f}_k^*) &\leq \int_0^\infty P\left\{E[\Delta(\mathbf{X}, \mathbf{f}_{k,n})|\mathcal{X}_n] - \Delta(\mathbf{f}_k^*) > t\right\} dt \\
&\leq u + 2\varepsilon + 2\left(|S_{k,\varepsilon}| + 1\right) \int_{u+2\varepsilon}^\infty e^{-2n((t/2)-\varepsilon)^2/D^4}\,dt \\
&\leq u + 2\varepsilon + 2\left(|S_{k,\varepsilon}| + 1\right) D^4 \cdot \frac{e^{-nu^2/(2D^4)}}{nu} \qquad (63) \\
&\leq \sqrt{\frac{2D^4 \log(|S_{k,\varepsilon}| + 1)}{n}} + 2\varepsilon + O(n^{-1/2}) \qquad (64)
\end{aligned}$$

where (63) follows from the inequality $\int_x^\infty e^{-t^2/2}\,dt < (1/x)e^{-x^2/2}$, for $x > 0$, and (64) follows by setting $u = \sqrt{\frac{2D^4 \log(|S_{k,\varepsilon}|+1)}{n}}$ where log denotes natural logarithm. The following lemma, which is proven below, demonstrates the existence of a suitable covering set $S_{k,\varepsilon}$.

Lemma 2 *For any $\varepsilon > 0$ there exists a finite collection of curves $S_{k,\varepsilon}$ in K such that*

$$\sup_{\mathbf{x} \in K} |\Delta(\mathbf{x}, \mathbf{f}) - \Delta(\mathbf{x}, \mathbf{f}')| \leq \varepsilon$$

and

$$|S_{k,\varepsilon}| \leq 2^{\frac{LD}{\varepsilon} + 3k + 1} V_d^{k+1} \left(\frac{D^2\sqrt{d}}{\varepsilon} + \sqrt{d}\right)^d \left(\frac{LD\sqrt{d}}{k\varepsilon} + 3\sqrt{d}\right)^{kd}$$

where V_d is the volume of the d-dimensional unit sphere and D is the diameter of K.

It is not hard to see that setting $\varepsilon = 1/k$ in Lemma 2 gives the upper bound

$$2D^4 \log(|S_{k,\varepsilon}| + 1) \leq kC(L, D, d)$$

where $C(L, D, d)$ does not depend on k. Combining this with (64) and the approximation bound given by (56) results in

$$\Delta(\mathbf{f}_{k,n}) - \Delta(\mathbf{f}^*) \leq \sqrt{\frac{kC(L, D, d)}{n}} + \frac{DL + 2}{k} + O(n^{-1/2}).$$

The rate at which $\Delta(\mathbf{f}_{k,n})$ approaches $\Delta(\mathbf{f}^*)$ is optimized by setting the number of segments k to be proportional to $n^{1/3}$. With this choice $J(\mathbf{f}_{k,n}) = \Delta(\mathbf{f}_{k,n}) - \Delta(\mathbf{f}^*)$ has the asymptotic convergence rate

$$J(\mathbf{f}_{k,n}) = O(n^{-1/3}),$$

and the proof of Theorem 2 is complete.

To show the bound (53), let $\delta \in (0, 1)$ and observe that by (62) we have

$$P\left\{E[\Delta(\mathbf{X}, \mathbf{f}_{k,n})|\mathcal{X}_n] - \Delta(\mathbf{f}_k^*) \leq t\right\} > 1 - \delta$$

whenever $t > 2\varepsilon$ and

$$\delta = 2\left(|S_{k,\varepsilon}| + 1\right) e^{-2n((t/2)-\varepsilon)^2/D^4}.$$

Solving this equation for t and letting $\varepsilon = 1/k$ as before, we obtain

$$\begin{aligned} t &= \sqrt{\frac{2D^4 \log\left(|S_{k,1/k}| + 1\right) - 2D^4 \log(\delta/2)}{n}} + \frac{2}{k} \\ &\leq \sqrt{\frac{kC(L,D,d) - 2D^4 \log(\delta/2)}{n}} + \frac{2}{k}. \end{aligned}$$

Therefore, with probability at least $1 - \delta$, we have

$$E\left[\Delta(\mathbf{X}, \mathbf{f}_{k,n}) | X_n\right] - \Delta(\mathbf{f}_k^*) \leq \sqrt{\frac{kC(L,D,d) - 2D^4 \log(\delta/2)}{n}} + \frac{2}{k}.$$

Combining this bound with the approximation bound $\Delta(\mathbf{f}_k^*) - \Delta(\mathbf{f}^*) \leq (DL)/k$ gives (53). □

Proof of Lemma 2 Consider a rectangular grid with side length $\delta > 0$ in $\mathbb{R}^d$. With each point $\mathbf{y}$ of this grid associate its Voronoi region (a hypercube of side length δ) defined as the set of points which are closer to $\mathbf{y}$ than to any other points of the grid. Let $K_\delta \subset K$ denote the collection of points of this grid which fall in K plus the projections of those points of the grid to K whose Voronoi regions have nonempty intersections with K. Then we clearly have

$$\max_{\mathbf{x} \in K} \min_{\mathbf{y} \in K_\delta} \|\mathbf{x} - \mathbf{y}\| \leq \frac{\sqrt{d}\delta}{2}. \tag{65}$$

Let $\delta = \varepsilon/(D\sqrt{d})$ and define $S_{k,\varepsilon}$ to be the family of all polygonal curves $\hat{\mathbf{f}}$ having $k+1$ vertices $\hat{\mathbf{y}}_0, \ldots, \hat{\mathbf{y}}_k \in K_\delta$ and satisfying the length constraint

$$l(\hat{\mathbf{f}}) \leq L + k\sqrt{d}\delta. \tag{66}$$

To see that $S_{k,\varepsilon}$ has the desired covering property, let $\mathbf{f} \in \S_k$ be arbitrary with vertices $\mathbf{y}_0, \ldots, \mathbf{y}_k$, choose $\hat{\mathbf{y}}_i \in K_\delta$ such that $\|\mathbf{y}_i - \hat{\mathbf{y}}_i\| \leq \sqrt{d}\delta/2$, $i = 0, \ldots, k$, and let $\hat{\mathbf{f}}$ be the polygonal curve with vertices $\hat{\mathbf{y}}_0, \ldots, \hat{\mathbf{y}}_k$. Since $\sum_i \|\mathbf{y}_i - \mathbf{y}_{i-1}\| \leq L$ by the definition of $\S_k$, the triangle inequality implies that $\hat{\mathbf{f}}$ satisfies (66) and thus $\hat{\mathbf{f}} \in \S_{k,\varepsilon}$. On the other hand, without loss of generality, assume that the line segment connecting $\mathbf{y}_{i-1}$ and $\mathbf{y}_i$ and the line segment connecting $\hat{\mathbf{y}}_{i-1}$ and $\hat{\mathbf{y}}_i$ are both linearly parameterized over $[0,1]$. Then

$$\begin{aligned} \max_{0 \leq t \leq 1} \|\mathbf{f}(t) - \hat{\mathbf{f}}(t)\| &= \max_{0 \leq t \leq 1} \|t\mathbf{y}_i + (1-t)\mathbf{y}_{i-1} - t\hat{\mathbf{y}}_i - (1-t)\hat{\mathbf{y}}_{i-1}\| \\ &\leq \max_{0 \leq t \leq 1} \left(t\|\mathbf{y}_i - \hat{\mathbf{y}}_i\| + (1-t)\|\mathbf{y}_{i-1} - \hat{\mathbf{y}}_{i-1}\|\right) \\ &\leq \frac{\sqrt{d}\delta}{2}. \end{aligned}$$

This shows that $\max\{\rho(\mathbf{f},\hat{\mathbf{f}}),\rho(\hat{\mathbf{f}},\mathbf{f})\} \leq \sqrt{d}\delta/2$. Then it follows from (54) that $S_{k,\varepsilon}$ is an ε-cover for $\S_k$ since for all $\mathbf{x} \in K$,

$$\begin{aligned} |\Delta(\mathbf{x},\mathbf{f}) - \Delta(\mathbf{x},\hat{\mathbf{f}})| &\leq 2D\max\{\rho(\mathbf{f},\hat{\mathbf{f}}),\rho(\hat{\mathbf{f}},\mathbf{f})\} \\ &\leq 2D\sqrt{d}\delta/2 = \varepsilon. \end{aligned}$$

Let L_i, $i = 1,\ldots,k$ denote the length of the ith segment of $\hat{\mathbf{f}}$ and let

$$\hat{L}_i = \left\lceil \frac{L_i}{\sqrt{d}\delta} \right\rceil \sqrt{d}\delta$$

where $\lceil x \rceil$ denotes the least integer not less than x. Fix the sequence $\hat{L}_1^k = \hat{L}_1,\ldots,\hat{L}_k$ and define $S_{k,\varepsilon}(\hat{L}_1^k) \subset \S_{k,\varepsilon}$ as the set of all $\hat{\mathbf{f}} \in S_{k,\varepsilon}$ whose segment lengths generate this particular sequence. To bound $|S_{k,\varepsilon}(\hat{L}_1^k)|$ note that the first vertex $\hat{\mathbf{y}}_0$ of an $\hat{\mathbf{f}} \in \S_{k,\varepsilon}(\hat{L}_1^k)$ can be any of the points in K_δ which contains as many points as there are Voronoi cells intersecting K. Since the diameter of K is D, there exists a sphere of radius $D+\sqrt{d}\delta$ which contains these Voronoi cells. Thus the cardinality of K_δ can be upper bounded as

$$|K_\delta| \leq V_d \left(\frac{D+\sqrt{d}\delta}{\delta} \right)^d$$

where V_d is the volume of the unit sphere in $\mathbb{R}^d$. Assume $\hat{\mathbf{y}}_0,\ldots,\hat{\mathbf{y}}_{i-1}$, $1 \leq i \leq k$ has been chosen. Since $\|\hat{\mathbf{y}}_i - \hat{\mathbf{y}}_{i-1}\| = L_i \leq \hat{L}_i$, there are no more than

$$V_d \left(\frac{L_i+\sqrt{d}\delta}{\delta} \right)^d \leq V_d \left(\frac{\hat{L}_i+\sqrt{d}\delta}{\delta} \right)^d$$

possibilities for choosing $\hat{\mathbf{y}}_i$. Therefore,

$$|S_{k,\varepsilon}(\hat{L}_1^k)| \leq V_d^{k+1} \left(\frac{D+\sqrt{d}\delta}{\delta} \right)^d \prod_{i=1}^k \left(\frac{\hat{L}_i+\sqrt{d}\delta}{\delta} \right)^d.$$

By (66) and the definition of $\hat{L}_i$, we have

$$\frac{1}{k}\sum_{i=1}^k (\hat{L}_i + \sqrt{d}\delta) \leq \frac{1}{k}\sum_{i=1}^k (L_i + 2\sqrt{d}\delta) \leq \frac{L}{k} + 3\sqrt{d}\delta. \tag{67}$$

Therefore, the arithmetic-geometric mean inequality implies that

$$\prod_{i=1}^k (\hat{L}_i + \sqrt{d}\delta) \leq \left(L/k + 3\sqrt{d}\delta \right)^k,$$

and thus

$$|S_{k,\varepsilon}(\hat{L}_1^k)| \leq V_d^{k+1} \left(\frac{D+\sqrt{d}\delta}{\delta} \right)^d \left(\frac{L}{k\delta} + 3\sqrt{d} \right)^{kd}.$$

On the other hand, by (67) we have $\sum_i \frac{\hat{L}_i}{\sqrt{d}\delta} \leq \frac{L}{\sqrt{d}\delta} + 2k$ and therefore the number of distinct sequences $\hat{L}_1^k$ is upper bounded by

$$\binom{\lceil \frac{L}{\sqrt{d}\delta} + 2k \rceil + k}{k} = \binom{\lceil \frac{L}{\sqrt{d}\delta} \rceil + 3k}{k} \leq 2^{\lceil \frac{L}{\sqrt{d}\delta} \rceil + 3k}.$$

Substituting $\delta = \varepsilon/(D\sqrt{d})$ we obtain

$$\begin{aligned} |S_{k,\varepsilon}| &= \sum_{\hat{L}_1^k} |S_{k,\varepsilon}(\hat{L}_1^k)| \\ &\leq 2^{\lceil \frac{LD}{\varepsilon} \rceil + 3k} V_d^{k+1} \left(\frac{D^2\sqrt{d}}{\varepsilon} + \sqrt{d} \right)^d \left(\frac{LD\sqrt{d}}{k\varepsilon} + 3\sqrt{d} \right)^{kd}. \end{aligned}$$

□