Efficient Regression in Metric Spaces via Approximate Lipschitz Extension*

Lee-Ad Gottlieb[1], Aryeh Kontorovich[2]**, and Robert Krauthgamer[3]***

[1] Ariel University
[2] Ben-Gurion University of the Negev
[3] Weizmann Institute of Science

Abstract. We present a framework for performing efficient regression in general metric spaces. Roughly speaking, our regressor predicts the value at a new point by computing a Lipschitz extension — the smoothest function consistent with the observed data — while performing an optimized structural risk minimization to avoid overfitting. The offline (learning) and online (inference) stages can be solved by convex programming, but this naive approach has runtime complexity $O(n^3)$, which is prohibitive for large datasets. We design instead an algorithm that is fast when the doubling dimension, which measures the "intrinsic" dimensionality of the metric space, is low. We make dual use of the doubling dimension: first, on the statistical front, to bound fat-shattering dimension of the class of Lipschitz functions (and obtain risk bounds); and second, on the computational front, to quickly compute a hypothesis function and a prediction based on Lipschitz extension. Our resulting regressor is both asymptotically strongly consistent and comes with finite-sample risk bounds, while making minimal structural and noise assumptions.

Keywords: metric space; regression; convex program

1 Introduction

The classical problem of estimating a continuous-valued function from noisy observations, known as *regression*, is of central importance in statical theory with a broad range of applications, see e.g. [BFOS84, Nad89, GKKW02]. When no structural assumptions concerning the target function are made, the regression problem is termed *nonparametric*. Informally, the main objective in the study of nonparametric regression is to understand the relationship between the regularity conditions that a function class might satisfy (e.g., Lipschitz or Hölder continuity, or sparsity in some representation) and the minimax risk convergence

* A full version appears as arXiv:1111.4470 [GKK11].

** This research was partially supported by the Israel Science Foundation (grant #1141/12) and the Lynne and William Frankel Center for Computer Science.

*** This work was supported in part by a US-Israel BSF grant #2010418, and by the Citi Foundation.

rates [Tsy04, Was06]. A further consideration is the computational efficiency of constructing the regression function.

The general (univariate) nonparametric regression problem may be stated as follows. Let $(\mathcal{X}, \rho)$ be a metric space, namely $\mathcal{X}$ is a set of points and ρ a distance function, and let $\mathcal{H}$ be a collection of functions ("hypotheses") $h : \mathcal{X} \to [0, 1]$. (Although in general, h is not explicitly restricted to have bounded range, typical assumptions on the diameter of $\mathcal{X}$ and the noise distribution amount to an effective truncation.) The space $\mathcal{X} \times [0, 1]$ is endowed with some fixed, unknown probability distribution μ, and the learner observes n iid draws $(X_i, Y_i) \sim \mu$. The learner then seeks to fit the observed data with some hypothesis $h \in \mathcal{H}$ so as to minimize the *risk*, usually defined as the expected loss $\mathbf{E}\,|h(X) - Y|^q$ for $(X, Y) \sim \mu$ and some $q \geq 1$.

Two limiting assumptions have traditionally been made when approaching this problem: (i) the space $\mathcal{X}$ is Euclidean and (ii) $Y_i = h^*(X_i) + \xi_i$, where h^* is the target function and ξ_i is an iid noise process, often taken to be Gaussian. Although our understanding of nonparametric regression under these assumptions is quite elaborate, little is known about nonparametric regression in the absence of either assumption.

The present work takes a step towards bridging this gap. Specifically, we consider nonparametric regression in an arbitrary metric space, while making no assumptions on the distribution of the data or the noise. Our results rely on the structure of the metric space only to the extent of assuming that the metric space has a low "intrinsic" dimensionality. The dimension in question is the *doubling dimension* of $\mathcal{X}$, denoted $\mathrm{ddim}(\mathcal{X})$, which was introduced by [GKL03] based on earlier work of [Cla99], and has been since utilized in several algorithmic contexts, including networking, combinatorial optimization, and similarity search, see e.g. [KSW09, KL04, BKL06, HM06, CG06, Cla06]. Following the work in [GKK10] on classification problems, our risk bounds and algorithmic runtime bounds are stated in terms of the doubling dimension of the ambient space and the Lipschitz constant of the regression hypothesis, although neither of these quantities need be known in advance.

Our results. We consider two kinds of risk: L_1 (mean absolute) and L_2 (mean square). More precisely, for $q \in \{1, 2\}$ we associate to each hypothesis $h \in \mathcal{H}$ the empirical L_q-risk

$$R_n(h) = R_n(h, q) = \frac{1}{n} \sum_{i=1}^{n} |h(X_i) - Y_i|^q \tag{1}$$

and the (expected) L_q-risk

$$R(h) = R(h, q) = \mathbf{E}\,|h(X) - Y|^q = \int_{\mathcal{X} \times [0,1]} |h(x) - y|^q \, \mu(dx, dy). \tag{2}$$

It is well-known that $h(x) = \mathbf{M}[Y \mid X = x]$ (where $\mathbf{M}$ is the median) minimizes $R(\cdot, 1)$ over all integrable $h \in [0, 1]^{\mathcal{X}}$ and $h(x) = \mathbf{E}[Y \mid X = x]$ minimizes

$R(\cdot, 2)$. However, these expressions are of little use as neither is computable without knowledge of μ. To circumvent this difficulty, we minimize the empirical L_q-risk and assert that the latter is a good approximation of the expected risk, provided $\mathcal{H}$ meets certain regularity conditions.

To this end, we define the following random variable, termed *uniform deviation*:

$$\Delta_n(\mathcal{H}) = \Delta_n(\mathcal{H}, q) = \sup_{h \in \mathcal{H}} |R_n(h) - R(h)| . \tag{3}$$

It is immediate that

$$R(h) \leq R_n(h) + \Delta_n(\mathcal{H}) \tag{4}$$

holds for all $h \in \mathcal{H}$ (i.e., the expected risk of any hypothesis does not exceed its empirical risk by much), and it can further be shown [BBL05] that $R(\hat{h}) \leq R(h^*) + 2\Delta_n(\mathcal{H})$, where $\hat{h} \in \mathcal{H}$ is a minimizer of the empirical risk and $h^* \in \mathcal{H}$ is a minimizer of the expected risk (i.e., the expected risk of $\hat{h}$ does not exceed the risk of the best admissible hypothesis by much).

Our contribution is twofold: statistical and computational. The algorithm in Theorem 3.1 computes an η-additive approximation to the empirical risk minimizer in time $\eta^{-O(\mathrm{ddim}(\mathcal{X}))} n \log^3 n$. This hypothesis can be evaluated on new points in time $\eta^{-O(\mathrm{ddim}(\mathcal{X}))} \log n$. The expected risk of this hypothesis decays as the empirical risk plus $1/\mathrm{poly}(n)$. Our bounds explicitly depend on the doubling dimension, but the latter may be efficiently estimated from the data, see e.g. [KL04, CG06, GK10, GKK13].

Related work. There are many excellent references for classical Euclidean nonparametric regression assuming iid noise, see for example [GKKW02, BFOS84, DGL96]. For metric regression, a simple risk bound follows from classic VC theory via the pseudo-dimension, see e.g. [Pol84, Vap95, Ney06]. However, the pseudo-dimension of many non-trivial function classes, including Lipschitz functions, grows linearly with the sample size, ultimately yielding a vacuous bound. An approach to nonparametric regression based on empirical risk minimization, though only for the Euclidean case, may already be found in [LZ95]; see the comprehensive historical overview therein. Indeed, Theorem 5.2 in [GKKW02] gives a kernel regressor for Lipschitz functions that achieves the minimax rate. Note however that (a) the setting is restricted to Euclidean spaces; and (b) the cost of evaluating the hypothesis at a new point grows linearly with the sample size (while our complexity is roughly logarithmic). As noted above, another feature of our approach is its ability to give efficiently computable finite-sample bounds, as opposed to the asymptotic convergence rates obtained in [GKKW02, LZ95] and elsewhere.

More recently, risk bounds in terms of doubling dimension and Lipschitz constant were given in [Kpo09], assuming an additive noise model, and hence these results are incomparable to ours; for instance, these risk bounds worsen with an increasingly smooth regression function. Following up, a regression technique

based on random partition trees was proposed in [KD11], based on mappings between Euclidean spaces and assuming an additive noise model. Another recent advance in nonparametric regression was Rodeo [LW08], which escapes the curse of dimensionality by adapting to the sparsity of the regression function.

Our work was inspired by the paper of von Luxburg and Bousquet [vLB04], who were apparently the first to make the connection between Lipschitz classifiers in metric spaces and large-margin hyperplanes in Banach spaces, thereby providing a novel generalization bound for nearest-neighbor classifiers. They developed a powerful statistical framework whose core idea may be summarized as follows: to predict the behavior at new points, find the smoothest function consistent with the training sample. Their work raises natural algorithmic questions like how to estimate the risk for a given input, how to perform model selection (Structural Risk Minimization) to avoid overfitting, and how to perform the learning and prediction quickly. Follow-up work [GKK10] leveraged the doubling dimension simultaneously for statistical and computational efficiency, to design an efficient classifier for doubling spaces. Its key feature is an efficient algorithm to find the optimal balance between the empirical risk and the penalty term for a given input. Minh and Hoffman [MH04] take the idea in [vLB04] in a more algebraic direction, establishing a representer theorem for Lipschitz functions on compact metric spaces.

2 Bounds on uniform deviation via fat shattering

In this section, we derive tail bounds on the uniform deviation $\Delta_n(\mathcal{H})$ defined in (3) in terms of the the smoothness properties of $\mathcal{H}$ and the doubling dimension of the underlying metric space $(\mathcal{X}, \rho)$.

2.1 Preliminaries

We rely on the powerful framework of fat-shattering dimension developed by [ABCH97], which requires us to incorporate the value of a hypothesis and the loss it incurs on a sample point into a single function. This is done by associating to any family of hypotheses $\mathcal{H}$ mapping $\mathcal{X} \mapsto [0,1]$, the induced family $\mathcal{F} = \mathcal{F}_{\mathcal{H}}^q$ of functions mapping $\mathcal{X} \times [0,1] \mapsto [0,1]$ as follows: for each $h \in \mathcal{H}$ the corresponding $f = f_h^q \in \mathcal{F}_{\mathcal{H}}^q$ is given by

$$f_h^q(x,y) = |h(x) - y|^q \,, \qquad q \in \{1,2\} \,. \tag{5}$$

In a slight abuse of notation, we define the uniform deviation of a class $\mathcal{F}$ of $[0,1]$-valued functions over $\mathcal{X} \times [0,1]$:

$$\Delta_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i, Y_i) - \mathbf{E} f(X,Y) \right| , \tag{6}$$

where the expectation is over μ, as in (2). Obviously, $\Delta_n(\mathcal{F}_{\mathcal{H}}^q) = \Delta_n(\mathcal{H}, q)$.

2.2 Basic generalization bounds

Let us write

$$\mathcal{H}_L = \left\{h \in [0,1]^{\mathcal{X}} : \|h\|_{\mathrm{Lip}} \leq L\right\} \tag{7}$$

to denote the collection of $[0,1]$-valued L-Lipschitz functions on $\mathcal{X}$. We proceed to bound the γ-fat-shattering dimension of $\mathcal{F}^q_{\mathcal{H}_L}$.

Theorem 2.1. *Let $\mathcal{H}_L$ be defined on a metric space $(\mathcal{X}, \rho)$, where $\mathrm{diam}(\mathcal{X}) = 1$. Then*

$$\mathrm{fat}_\gamma(\mathcal{F}^q_{\mathcal{H}_L}) \leq \left(1 + \frac{1}{\gamma^{(q+1)/2}}\right)\left(\frac{L}{\gamma^{(q+1)/2}}\right)^{\mathrm{ddim}(\mathcal{X})+1}$$

holds for $q \in \{1,2\}$ and all $0 < \gamma \leq \frac{1}{2}$.

Proof. (Sketch) Fix a $\gamma > 0$ and recall what it means for $\mathcal{F}^q_{\mathcal{H}_L}$ to γ-shatter a set

$$S = (T, Z) = \{(t, z) : t \in \mathcal{X}, z \in [0,1]\}$$

(where $T \in \mathcal{X}^{|S|}$ and $Z \in [0,1]^{|S|}$): there exists some function $r \in \mathbb{R}^S$ such that for each label assignment $b \in \{-1,1\}^S$ there is an $f \in \mathcal{F}^q_{\mathcal{H}_L}$ satisfying $b(s)(f(s) - r(s)) \geq \gamma$ for all $s \in S$.

Put $K = \left\lceil \gamma^{-(q+1)/2} \right\rceil$ and define the map $\pi : S \to \{0, 1, \ldots, K\}$ by

$$\pi(s) = \pi(t, z) = \lfloor Kz \rfloor .$$

Thus, we may view S as being partitioned into $K+1$ buckets:

$$S = \bigcup_{k=0}^{K} \pi^{-1}(k). \tag{8}$$

Consider two points, $s = (t, z)$ and $s' = (t', z')$, belonging to some fixed bucket $\pi^{-1}(k)$. By construction, the following hold:

(i) $|z - z'| \leq K^{-1} \leq \gamma^{(q+1)/2}$
(ii) since $\mathcal{F}^q_{\mathcal{H}_L}$ γ-shatters S (and recalling (5)), there is an $h \in \mathcal{H}_L$ satisfying $|h(t) - z|^q \leq r - \gamma$ and $|h(t') - z'|^q \geq r' + \gamma$ for some $\gamma \leq r \leq r' < 1 - \gamma$.

Conditions (i) and (ii) imply that

$$|h(t) - h(t')| \geq (r' + \gamma)^{1/q} - (r - \gamma)^{1/q} - |z - z'| \geq \gamma^{(q+1)/2}. \tag{9}$$

The fact that h is L-Lipschitz implies that $\rho(t, t') \geq |h(t) - h(t')|/L \geq \gamma^{(q+1)/2}/L$ and hence

$$\left|\pi^{-1}(k)\right| \leq \left(\frac{L}{\gamma^{(q+1)/2}}\right)^{\mathrm{ddim}(\mathcal{X})+1} \tag{10}$$

for each $k \in \left\{0, 1, \ldots, \left\lceil \gamma^{-(q+1)/2} \right\rceil\right\}$. Together (8) and (10) yield our desired bound on $|S|$, and hence on the fat shattering dimension of $\mathcal{F}^q_{\mathcal{H}_L}$. □

The following generalization bound, implicit in [ABCH97], establishes the learnability of continuous-valued functions in terms of their fat-shattering dimension.

Theorem 2.2. *Let $\mathcal{F}$ be any admissible function class mapping $\mathcal{X} \times [0,1]$ to $[0,1]$ and define $\Delta_n(\mathcal{F})$ as in (6). Then for all $0 < \varepsilon < 1$ and all $n \geq 2/\varepsilon^2$,*

$$P(\Delta_n(\mathcal{F}) > \varepsilon) \leq 24n \left(\frac{288n}{\varepsilon^2}\right)^{d \log(24en/\varepsilon)} \exp(-\varepsilon^2 n/36)$$

where $d = \mathrm{fat}_{\varepsilon/24}(\mathcal{F})$.

Corollary 2.1. *Fix an $1 > \varepsilon > 0$ and $q \in \{1, 2\}$. Let $\mathcal{H}_L$ be defined on a metric space $(\mathcal{X}, \rho)$ and recall the definition of $\Delta_n(\mathcal{H}_L, q)$ in (3). Then for all $n \geq 2/\varepsilon^2$,*

$$P(\Delta_n(\mathcal{H}_L, q) > \varepsilon) \leq 24n \left(\frac{288n}{\varepsilon^2}\right)^{d \log(24en/\varepsilon)} \exp(-\varepsilon^2 n/36) \tag{11}$$

where

$$d = \left(1 + \frac{1}{(\varepsilon/24)^{(q+1)/2}}\right) \left(\frac{L}{(\varepsilon/24)^{(q+1)/2}}\right)^{\mathrm{ddim}(\mathcal{X})+1}.$$

We can conclude from Corollary 2.1 that there exists $\epsilon(n, L, \delta)$ such that with probability at least $1 - \delta$,

$$\Delta_n(\mathcal{H}_L, q) \leq \epsilon(n, L, \delta), \tag{12}$$

and by essentially inverting (11), we have

$$\epsilon(n, L, \delta) \leq O\left(\max\left\{\sqrt{\frac{\log(n/\delta)}{n}}, \left(\frac{L^{\mathrm{ddim}(\mathcal{X})+1}}{n} \log^2 n\right)^{\frac{1}{2+\frac{q+1}{2}(\mathrm{ddim}(\mathcal{X})+1)}}\right\}\right) \tag{13}$$

(For simplicity, the dependence of $\epsilon(\cdot)$ on $\mathrm{ddim}(\mathcal{X})$ is suppressed.) This implies via (4) that

$$R(h) \leq R_n(h) + \epsilon(n, L, \delta)$$

uniformly for all $h \in \mathcal{H}_L$ with high probability.

2.3 Simultaneous bounds for multiple Lipschitz constants

So far, we have established the following. Let $(\mathcal{X}, \rho)$ be a doubling metric space and $\mathcal{H}_L$ a collection of L-Lipschitz $[0,1]$-valued functions on $\mathcal{X}$. Then Corollary 2.1 guarantees that for all $\varepsilon, \delta \in (0,1)$ and $n \geq n_0(\varepsilon, \delta, L, \mathrm{ddim}(\mathcal{X}))$, we have

$$P(\Delta_n(\mathcal{H}_L) > \varepsilon) \leq \delta, \tag{14}$$

where $\Delta_n(\mathcal{H}_L)$ is the uniform deviation defined in (3). Since our computational approach in Section 3 requires optimizing over Lipschitz constants, we will need a bound such as (14) that holds for many function classes of varying smoothness simultaneously. This is easily accomplished by stratifying the confidence parameter δ, as in [SBWA98]. We will need the following theorem:

Theorem 2.3. *Let*

$$\mathcal{H}^{(1)} \subset \mathcal{H}^{(2)} \subset \dots$$

be a sequence of function classes taking $\mathcal{X}$ *to* $[0,1]$ *and let* $p_k \in [0,1]$, $k = 1, 2, \dots$, *be a sequence summing to* 1. *Suppose that* $\epsilon : \mathbb{N} \times \mathbb{N} \times (0,1) \to [0,1]$ *is a function such that for each* $k \in \mathbb{N}$, *with probability at least* $1 - \delta$, *we have*

$$\Delta_n^q(\mathcal{H}^{(k)}) \leq \epsilon(n, k, \delta).$$

Then, whenever some $h \in \bigcup_{k \in \mathbb{N}} [\mathcal{H}^{(k)}]_\eta$ *achieves empirical risk* $R_n(h)$ *on a sample of size* n, *we have that with probability at least* $1 - \delta$,

$$R(h) \leq R_n(h) + \epsilon(n, k, \delta p_k) \quad \forall k. \tag{15}$$

Proof. An immediate consequence of the union bound. □

The structural risk minimization principle implied by Theorem 2.3 amounts to the following model selection criterion: choose an $h \in \mathcal{H}^{(k)}$ for which the right-hand side of (15) is minimized.

In applying Theorem 2.3 to Lipschitz classifiers in Section 3 below, we impose a discretization on the Lipschitz constant L to be multiples of $\frac{\eta}{24q}$. Formally, we consider the stratification $\mathcal{H}^{(k)} = \mathcal{H}_{L_k}$,

$$\mathcal{H}_{L_1} \subset \mathcal{H}_{L_2} \subset \dots,$$

where $L_k = k\eta$ with corresponding $p_k = 2^{-k}$ for $k = 1, 2, \dots$. This means that whenever we need a hypothesis that is an L-Lipschitz regression function, we may take $k = \lceil L\eta \rceil$ and use $\epsilon(n, k, \delta 2^{-k})$ as the generalization error bound. Note that all possible values of L are within a factor of 2 of the discretized sequence L_k.

3 Structural risk minimization

In this section, we address the problem of efficient model selection when given n observed samples. The algorithm described below computes a hypothesis that approximately attains the minimum risk over all hypotheses. Since our approximate Lipschitz extension algorithm will evaluate hypotheses up to an additive error, we define an η-perturbation $[\mathcal{H}]_\eta$ of a given hypothesis class $\mathcal{H}$ by

$$[\mathcal{H}]_\eta = \left\{ h' \in \mathbb{R}^{\mathcal{X}} : \exists h \in \mathcal{H} \text{ s.t. } \|h - h'\|_\infty \leq \eta \right\}. \tag{16}$$

Recall the risk bound achieved as a consequence of Theorem 2.3. In the full paper [GKK11], we extend this result to perturbations, showing that whenever some $h \in \bigcup_{k \in \mathbb{N}} \left[\mathcal{H}^{(k)}\right]_\eta$ achieves empirical risk $R_n(h)$ on a sample of size n, we have the following bound on $R(h)$, the true risk of h:

$$R(h) \leq R_n(h) + \epsilon(n, k, \delta p_k) + 24q\eta, \tag{17}$$

with probability at least $1-\delta$ (where the diameter of the point set has been taken as 1, and $\epsilon(n,k,\delta p_k) \geq \sqrt{2/n}$ is the minimum value of ϵ for which the right-hand side of (11) is at most δ). In the rest of this section, we devise an algorithm that computes a hypothesis that approximately minimizes our bound from (17) on the true risk, denoted henceforth

$$\tilde{R}_\eta(h) = R_n(h) + \epsilon(n,k,\delta p_k) + 24q\eta.$$

Notice that on the right-hand side, the first two terms depend on L, but only the first term depends on the choice of h, and only the third term depends on η.

Theorem 3.1. *Let (X_i, Y_i) for $i = 1,\ldots,n$ be an iid sample drawn from μ, let $\eta \in (0, \frac{1}{4})$, and let h^* be a hypothesis that minimizes $\tilde{R}_\eta(h)$ over all $h \in \bigcup_{k\in\mathbb{N}} \left[\mathcal{H}^{(k)}\right]_\eta$. There is an algorithm that, given the n samples and η as input, computes in time $\eta^{-O(\mathrm{ddim}(\mathcal{X}))} n \log^3 n$ a hypothesis $h' \in \bigcup_{k\in\mathbb{N}} \left[\mathcal{H}^{(k)}\right]_\eta$ with*

$$\tilde{R}_\eta(h') \leq 2\tilde{R}_\eta(h^*). \tag{18}$$

Remark. We show in Theorem 4.1 how to quickly evaluate the hypothesis h' on new points.

The rest of Section 3 is devoted to describing an algorithm that realizes the bounds of Theorem 3.1 for $q = 1$ (Sections 3.1 and 3.2) and $q = 2$ (Section 3.3). In proving the theorem, we will find it convenient to compare the output h' to a hypothesis $\bar{h}$ that is smooth (i.e. Lipschitz but unperturbed). Indeed, let h^* be as in the theorem, and $\bar{h} \in \bigcup_{k\in\mathbb{N}} \mathcal{H}^{(k)}$ be a hypothesis that minimizes $\tilde{R}_\eta(\bar{h})$. Then $R_n(h^*) \leq R_n(\bar{h}) \leq R_n(h^*) + \eta$, and we get $\tilde{R}_\eta(h^*) \leq \tilde{R}_\eta(\bar{h}) \leq \tilde{R}_\eta(h^*) + \eta$. Accordingly, the analysis below will actually prove that $\tilde{R}_\eta(h') \leq 2\tilde{R}_\eta(\bar{h}) - 2\eta$, and then (18) will follow easily, essentially increasing the additive error by 2η. Moreover, once (18) is proved, we can use the above to conclude that $\tilde{R}_\eta(h') \leq 2\tilde{R}_0(\bar{h}) + O(\eta)$, which compares the risk bound of our algorithm's output h' to what we could possibly get using smooth hypotheses.

In the rest of this section we consider the n observed samples as fixed values, given as input to the algorithm, so we will write x_i instead of X_i.

3.1 Motivation and construction

Suppose that the Lipschitz constant of an optimal *unperturbed* hypothesis $\bar{h}$ were known to be $L = \bar{L}$. Then $\epsilon(n,k,\delta p_k)$ is fixed, and the problem of computing both $\bar{h}$ and its empirical risk $R_n(\bar{h})$ can be described as the following optimization program with variables $f(x_i)$ for $i \in [n]$ to represent the assignments $h(x_i)$. Note it is a Linear Program (LP) when $q = 1$ and a quadratic program when $q = 2$.

$$\begin{array}{|ll|}\hline \text{Minimize} & \sum_{i\in[n]} |y_i - f(x_i)|^q \\ \text{subject to} & |f(x_i) - f(x_j)| \leq L \cdot \rho(x_i,x_j)\ \forall i,j \in [n] \\ & 0 \leq f(x_i) \leq 1 \qquad\qquad \forall i \in [n] \\ \hline \end{array} \tag{19}$$

It follows that $\bar{h}$ could be computed by first deriving $\bar{L}$, and then solving the above program. However, it seems that computing these exactly is an expensive computation. This motivates our search for an approximate solution to risk minimization. We first derive a target Lipschitz constant L' that "approximates" $\bar{L}$, in the sense that there exists an h' with Lipschitz constant L' which minimizes the objective $\max\{R_n(h'), \epsilon(n,k,\delta p_k)\}$. Notice that $R_n(h')$ may be computed by solving LP (19) using the given value L' for L. We wish to find such L' via a binary search procedure, which requires a method to determine whether a candidate L satisfies $L \leq L'$, but since our objective need not be a monotone function of L, we cannot rely on the value of the objective at the candidate L. Instead, recall that the empirical risk term $R_n(h')$ is monotonically non-increasing, and the penalty term $\epsilon(n,k,\delta p_k)$ is monotonically non-decreasing, and therefore we can take L' to be the minimum value L for which $R_n(h') \leq \epsilon(n,k,\delta p_k)$ (notice that both terms are right-continuous in L). Our binary search procedure can thus determine whether a candidate L satisfies $L \leq L'$ by checking instead whether $R_n(h') \leq \epsilon(n,k,\delta p_k)$.

Were the binary search on L to be carried out indefinitely (that is, with infinite precision), it would yield L' and a smooth hypothesis h' satisfying $\tilde{R}_\eta(h') \leq 2\tilde{R}_\eta(\bar{h})$, where the factor 2 originates from the gap between maximum and summation. In fact, a slightly stronger bound holds:

$$\tilde{R}_\eta(h') - 24q\eta \leq 2\max\{R_n(h'), \epsilon(n,k,\delta p_k)\} \leq 2\big(R_n(\bar{h}) + \epsilon(n,k,\delta p_k)\big) \leq 2\big(\tilde{R}_\eta(\bar{h}) - 24q\eta\big).$$

(In our actual LP solver below, h' will not be necessarily smooth, but rather a perturbation of a smooth hypothesis.) However, to obtain a tractable runtime, we fix an additive precision of η to the Lipschitz constant, and restrict the target Lipschitz constant to be a multiple of η. Notice that $\tilde{R}_\eta(\bar{h}) \leq 2$ for sufficiently large n (since this bound can even be achieved by a hypothesis with Lipschitz constant 0), so by (13) it must be that $\bar{L} \leq n^{O(1)}$, since $\bar{L}$ is the optimal Lipschitz constant. It follows that the binary search will consider only $O(\log(n/\eta))$ candidate values for L'.

To bound the effect of discretizing the target L' to multiples of η, we shall show the existence of a hypothesis $\hat{h}$ that has Lipschitz constant $\hat{L} \leq \max\{\bar{L} - \eta, 0\}$ and satisfies $\tilde{R}_\eta(\hat{h}) \leq \tilde{R}_\eta(\bar{h}) + \eta$. To see this, assume by translation that the minimum and maximum values assigned by $\bar{h}$ are, respectively 0 and $a \leq 1$. Thus, its Lipschitz constant is $\bar{L} \geq a$ (recall we normalized $\mathrm{diam}(\mathcal{X}) = 1$). Assuming first the case $a \geq \eta$, we can set $\hat{h}(x) = (1 - \frac{\eta}{a}) \cdot \bar{h}(x)$, and it is easy to verify that its Lipschitz constant is at most $(1 - \frac{\eta}{a})\bar{L} \leq \bar{L} - \eta$, and $\tilde{R}_\eta(\hat{h}) \leq \tilde{R}_\eta(\bar{h}) + \eta$. The case $a < \eta$ is even easier, as now there is trivially a function $\hat{h}$ with Lipschitz constant 0 and $\tilde{R}_\eta(\hat{h}) \leq \tilde{R}_\eta(\bar{h}) + \eta$. It follows that when the binary search is analyzed using this $\hat{h}$ instead of $\bar{h}$, we actually get

$$\tilde{R}_\eta(h') \leq 2\tilde{R}_\eta(\hat{h}) - 24q\eta \leq 2\tilde{R}_\eta(\bar{h}) - 22q\eta \leq 2\tilde{R}_\eta(h^*) - 20q\eta.$$

It now remains to show that given L', program (19) may be solved quickly (within certain accuracy), which we do in Sections 3.2 and 3.3.

3.2 Solving the linear program

We show how to solve the linear program, given the target Lipschitz constant L'.

Fast LP-solver framework. To solve the linear program, we utilize the framework presented by Young [You01] for LPs of the following form: Given non-negative matrices P, C, vectors p, c and precision $\beta > 0$, find a non-negative vector x such that $Px \leq p$ and $Cx \geq c$. Young shows that if there exists a feasible solution to the input instance, then a solution to a relaxation of the input program (specifically, $Px \leq (1+\beta)p$ and $Cx \geq c$) can be found in time $O(md(\log m)/\beta^2)$, where m is the number of constraints in the program and d is the maximum number of constraints in which a single variable may appear.

In utilizing this framework for our problem, we encounter a difficulty that both the input matrices and output vector must be non-negative, while our LP (19) has difference constraints. To bypass this limitation, for each LP variable $f(x_i)$ we introduce a new variable $\tilde{x}_i$ and two new constraints:

$$f(x_i) + \tilde{x}_i \leq 1$$
$$f(x_i) + \tilde{x}_i \geq 1$$

By the guarantees of the LP solver, we have that in the returned solution $1 - f(x_i) \leq \tilde{x}_i \leq 1 - f(x_i) + \beta$ and $\tilde{x}_i \geq 0$. This technique allows us to introduce negated variables $-f(x_i)$ into the linear program, at the loss of additive precision.

Reduced constraints. A central difficulty in obtaining a near-linear runtime for the above linear program is that the number of constraints in LP (19) is $\Theta(n^2)$. We show how to reduce the number of constraints to near-linear in n, namely, $\eta^{-O(\mathrm{ddim}(\mathcal{X}))}n$. We will further guarantee that each of the n variables $f(x_i)$ appears in only $\eta^{-O(\mathrm{ddim}(\mathcal{X}))}$ constraints. Both these properties will prove useful for solving the program quickly.

Recall that the purpose of the $\Theta(n^2)$ constraints is solely to ensure that the target Lipschitz constant is not violated between any pair of points. We will show below that this property can be approximately maintained with many fewer constraints: The spanner described in our full paper [GKK11], has stretch $1 + \delta$, degree $\delta^{-O(\mathrm{ddim}(\mathcal{X}))}$ and hop-diameter $c' \log n$ for some constant $c' > 0$, that can be computed quickly. Build this spanner for the observed sample points $\{x_i : i \in [n]\}$ with stretch $1+\eta$ (i.e., set $\delta = \eta$) and retain a constraint in LP (19) if and only if its two variables correspond to two nodes that are connected in the spanner. It follows from the bounded degree of the spanner that each variable appears in $\eta^{-O(\mathrm{ddim}(\mathcal{X}))}$ constraints, which implies that there are $\eta^{-O(\mathrm{ddim}(\mathcal{X}))}n$ total constraints.

Modifying remaining constraints. Each spanner-edge constraint $|f(x_i) - f(x_j)| \leq L' \cdot \rho(x_i, x_j)$ is replaced by a set of two constraints

$$f(x_i) + \tilde{x}_j \leq 1 + L' \cdot \rho(x_i, x_j)$$
$$f(x_j) + \tilde{x}_i \leq 1 + L' \cdot \rho(x_i, x_j)$$

By the guarantees of the LP solver we have that in the returned solution, each spanner edge constraint will satisfy

$$\begin{aligned}|f(x_i) - f(x_j)| &\leq -1 + (1+\beta)[1 + L' \cdot \rho(x_i, x_j)] \\ &= \beta + (1+\beta)L' \cdot \rho(x_i, x_j).\end{aligned}$$

Now consider the Lipschitz condition for two points not connected by a spanner edge: Let $x_1, \ldots, x_{k+1}$ be a $(1+\eta)$-stretch ($k \leq c' \log n$)-hop spanner path connecting points $x = x_1$ and $x' = x_{k+1}$. Then the spanner stretch guarantees that

$$\begin{aligned}|f(x) - f(x')| &\leq \textstyle\sum_{i=1}^{k} [\beta + (1+\beta)L' \cdot \rho(x_i, x_{i+1})] \\ &\leq \beta c' \log n + (1+\beta)L' \cdot (1+\eta)\rho(x, x').\end{aligned}$$

Choosing $\beta = \frac{\eta^2}{24qc' \log n}$, and noting that $(1+\beta)(1+\eta) < (1+2\eta)$, we have that for all point pairs

$$|f(x) - f(x')| < \tfrac{\eta^2}{24q} + (1+2\eta)L' \cdot \rho(x, x').$$

We claim that the above inequality ensures that the computed hypothesis h' (represented by variables $f(x_i)$ above) is a 6η-perturbation of some hypothesis with Lipschitz constant L'. To prove this, first note that if $L' = 0$, then the statement follows trivially. Assume then that (by the discretization of L'), $L' \geq \eta$. Now note that a hypothesis with Lipschitz constant $(1+3\eta)L'$ is a 3η-perturbation of some hypothesis with Lipschitz constant L'. (This follows easily by scaling down this hypothesis by a factor of $(1+3\eta)$, and recalling that all values are in the range $[0, 1]$.) Hence, it suffices to show that the computed hypothesis h' is a 3η-perturbation of some hypothesis $\tilde{h}$ with Lipschitz constant $(1+3\eta)L'$. We can construct $\tilde{h}$ as follows: Extract from the sample points $S = \{x_i\}_{i \in [n]}$ a (η/L')-net N, then for every net-point $z \in N$ set $\tilde{h}(z) = h'(z)$, and extend this function $\tilde{h}$ from N to all of S without increasing Lipschitz constant by using the McShane-Whitney extension theorem [McS34, Whi34] for real-valued functions. Observe that for every two net-points $z \neq z' \in N$,

$$|\tilde{h}(z) - \tilde{h}(z')| \leq \frac{\eta^2}{24q} + (1+2\eta)L' \cdot \rho(z, z') < (1+3\eta)L' \cdot \rho(z, z').$$

It follows that $\tilde{h}$ (defined on all of S) has Lipschitz constant $\tilde{L} \leq 1 + 3\eta$. Now, consider any point $x \in S$ and its closest net-point $z \in N$; then $\rho(x, z) \leq \eta/L'$. Using the fact $\tilde{h}(z) = h'(z)$, we have that $|h'(x) - \tilde{h}(x)| \leq |h'(x) - h'(z)| + |\tilde{h}(z) - \tilde{h}(x)| \leq \left[\frac{\eta^2}{24q} + (1+2\eta)L' \cdot \rho(x, z)\right] + (1+3\eta)L' \cdot \rho(x, y) \leq \frac{\eta^2}{24q} + (2+5\eta)\eta \leq 3\eta$. We conclude that h' is 3η-perturbation of $\tilde{h}$, and a 6η-perturbation of some hypothesis with Lipschitz constant L'.

Objective function. We now turn to the objective function $\frac{1}{n}\sum_i |y_i - f(x_i)|$. We use the same technique as above for handling difference constraints: For each

term $|y_i - f(x_i)|$ in the objective function we introduce the variable w_i and the constraint

$$f(x_i) + w_i \geq y_i$$

Note that the solver imposes the constraint that $w_i \geq 0$, so we have that $w_i \geq \max\{0, y_i - f(x_i)\}$. Now consider the term $f(x_i) + 2w_i$, and note that the minimum feasible value of this term in the solution of the linear program is exactly equal to $y_i + |y_i - f(x_i)|$: If $f(x_i) \geq y_i$ then the minimum feasible value of w_i is 0, which yields $f(x_i) + 2w_i = f(x_i) = y_i + (f(x_i) - y_i) = y_i + |y_i - f(x_i)|$. Otherwise we have that $f(x_i) < y_i$, so the minimum feasible value of w_i is $y_i - f(x_i)$, which yields $f(x_i) + 2w_i = 2y_i - f(x_i) = y_i + |y_i - f(x_i)|$.

The objective function is then replaced by the constraint

$$\tfrac{1}{n} \textstyle\sum_i (f(x_i) + 2w_i) \leq r,$$

which by the above discussion is equal to $\frac{1}{n} \sum_i (y_i + |y_i - f(x_i)|) \leq r$, and hence is a direct bound on the empirical error of the hypothesis. We choose bound r via binary search: Recalling that $\tilde{R}_n(h') \leq 1$ (since even a hypothesis with Lipschitz constant 0 can achieve this bound), we may set $r \leq 1$. By discretizing r in multiples of η (similar to what was done for L'), we have that the binary search will consider only $O(\log \eta^{-1})$ guesses for r. Note that for guess r', the solver guarantees only that the returned sum is less than $(1 + \beta)r' \leq r' + \beta < r' + \eta$. It follows that the discretization of r and its solver relaxation of r introduce, together, at most an additive error of 2η in the LP objective, i.e., in $R_n(h')$ and in $\tilde{R}_\eta(h')$.

Correctness and runtime analysis. The fast LP solver ensures that h' computed by the above-described algorithm is a 6η-perturbation of a hypothesis with Lipschitz constant L'. As for $\tilde{R}(h')$, which we wanted to minimize, an additive error of 2η is incurred by comparing h' to $\bar{h}$ instead of to h^*, another additive error of 2η arises from discretizing $\bar{L}$ into L' (i.e., comparing to $\hat{h}$ instead of $\bar{h}$), and another additive error 4η introduced through the discretization of r and its solver relaxation. Overall, the algorithm above computes a hypothesis $h' \in \bigcup_{k \in \mathbb{N}} \left[\mathcal{H}^{(k)}\right]_{6\eta}$ with $\tilde{R}_\eta(h') \leq 2\tilde{R}_\eta(h^*) - 16\eta$. The parameters in Theorem 3.1 are achieved by scaling down η to $\frac{\eta}{6}$ and the simple manipulation $\tilde{R}_{\eta/6}(h) = \tilde{R}_\eta(h) - 20q\eta$.

Finally, we turn to analyze the algorithmic runtime. The spanner may be constructed in time $O(\eta^{-O(\mathrm{ddim}(\mathcal{X}))} n \log n)$. Young's LP solver [You01] is invoked $O(\log \frac{n}{\eta} \log \frac{1}{\eta})$ times, where the $\log \frac{n}{\eta}$ term is due to the binary search for L', and the $\log \frac{1}{\eta}$ term is due to the binary search for r. To determine the runtime per invocation, recall that each variable of the program appears in $d = \eta^{-O(\mathrm{ddim}(\mathcal{X}))}$ constraints, implying that there exist $m = \eta^{-O(\mathrm{ddim}(\mathcal{X}))} n$ total constraints. Since we set $\beta = O(\eta^2 / \log n)$, we have that each call to the solver takes time $O(md(\log m)/\beta^2) \leq \eta^{-O(\mathrm{ddim}(\mathcal{X}))} n \log^2 n$, for a total runtime of $\eta^{-O(\mathrm{ddim}(\mathcal{X}))} n \log^2 n \log \frac{n}{\eta} \log \frac{1}{\eta} \leq \eta^{-O(\mathrm{ddim}(\mathcal{X}))} n \log^3 n$. This completes the proof of Theorem 3.1 for $q = 1$.

3.3 Solving the quadratic program

Above, we considered the case when the loss function is linear. Here we modify the objective function construction to cover the case when the loss function is quadratic, that is $\frac{1}{n}\sum_i |y_i - f(x_i)|^2$. We then use the LP solver to solve our quadratic program. (Note that the spanner-edge construction above remains as before, and only the objective function construction is modified.)

Let us first redefine w_i by the constraints

$$f(x_i) + w_i \leq 1$$
$$f(x_i) + w_i \geq 1$$

It follows from the guarantees of the LP solver that in the returned solution, $1 - f(x_i) \leq w_i \leq 1 - f(x_i) + \beta$ and $w_i \geq 0$.

Now note that a quadratic inequality $v \geq x^2$ can be approximated for $x \in [0, 1]$ by a set of linear inequalities of the form

$$v \geq 2(j\eta)x - (j\eta)^2$$

for $0 \leq j \leq \frac{1}{\eta}$; these are just a collection of tangent lines to the quadratic function. Note that the slope of the quadratic function in the stipulated range is at most 2, so this approximation introduces an additive error of at most 2η.

Since $|y_i - f(x_i)|^2$ takes values in the range $[0, 1]$, we will consider an equation set of the form

$$v_i \geq 2(j\eta)|y_i - f(x_i)| - (j\eta)^2 + 2\eta$$

which satisfies that the minimum feasible value of v_i is in the range $[|y_i - f(x_i)|^2, |y_i - f(x_i)|^2 + 2\eta]$. It remains to model these difference constraints in the LP framework: When $f(x_i) \leq y_i$, the equation set

$$v_i + 2(j\eta)f(x_i) \geq 2(j\eta)y_i - (j\eta)^2 + 2\eta$$

exactly models the above constraints. When $f(x_i) > y_i$, the lower bound of this set may not be tight, and instead the equation set

$$v_i + 2(j\eta)w_i \geq -2(j\eta)y_i - (j\eta)^2 + 2\eta + 2(j\eta)(1 + \beta)$$

models the above constraints, though possibly increasing the value of v_i by $2(j\eta)\beta < \eta$. (Note that when $f(x_i) < y_i$, the lower bound of the second equation set may not be tight, so the first equation set is necessary. Also, note that whenever the right hand side of an equation is negative, the equation is vacuous and may be omitted.)

The objective function is then replaced by the inequality

$$\frac{1}{n}\sum_i v_i \leq r,$$

where r is chosen by binary search as above.

Turning to the runtime analysis, the replacement of a constraint by $O(1/\eta)$ new constraints does not change the asymptotic runtime. For the analysis of the

approximation error, first note that a solution to this program is a feasible solution to the original quadratic program. Further, given a solution to the original quadratic program, a feasible solution to the above program can be found by perturbing the quadratic program solution by at most 3η (since additive terms of 2η and η are lost in the above construction). The proof of Theorem 3.1 for $q = 2$ follows by an appropriate scaling of η.

4 Approximate Lipschitz extension

In this section, we show how to evaluate our hypothesis on a new point. More precisely, given a hypothesis function $f : S \to [0, 1]$, we wish to evaluate a minimum Lipschitz extension of f on a new point $x \notin S$. That is, denoting $S = \{x_1, \ldots, x_n\}$, we wish to return a value $y = f(x)$ that minimizes $\max_i\{\frac{|y-f(x_i)|}{\rho(x,x_i)}\}$. Necessarily, this value is not greater than the Lipschitz constant of the classifier, meaning that the extension of f to the new point does not increase the Lipschitz constant of f and so Theorem 2.3 holds for the single new point. (By this local regression analysis, it is not necessary for newly evaluated points to have low Lipschitz constant with respect to each other, since Theorem 2.3 holds for each point individually.)

First note that the Lipschitz extension label y of $x \notin S$ will be determined by two points of S. That is, there are two points $x_i, x_j \in S$, one with label greater than y and one with a label less than y, such that the Lipschitz constant of (x, y) relative to each of these points (that is, $L = \frac{f(x_i)-y}{\rho(x,x_i)} = \frac{y-f(x_j)}{\rho(x,x_j)}$) is maximum over the Lipschitz constant of (x, y) relative to any point in S. Hence, y cannot be increased or decreased without increasing the Lipschitz constant with respect to one of these points.

Note then that an exact Lipschitz extension may be derived in $\Theta(n^2)$ time in brute-force fashion, by enumerating all point pairs in S, calculating the optimal Lipschitz extension for x with respect to each pair alone, and then choosing the candidate value for y with the highest Lipschitz constant. However, we demonstrate that an approximate solution to the Lipschitz extension problem can be derived more efficiently.

Theorem 4.1. *An η-additive approximation to the Lipschitz extension problem can be computed in time $\eta^{-O(\mathrm{ddim}(\mathcal{X}))} \log n$.*

Proof. The algorithm is as follows: Round up all labels $f(x_i)$ to the nearest term $j\eta/2$ (for any integer $0 \leq j \leq 2/\eta$), and call the new label function $\tilde{f}$. We seek the value of $\tilde{f}(x)$, the optimal Lipschitz extension value for x for the new function $\tilde{f}$. Trivially, $f(x) \leq \tilde{f}(x) \leq f(x) + \eta/2$. Now, if we were given for each j the point with label $j\eta/2$ that is the nearest neighbor of x (among all points with this label), then we could run the brute-force algorithm described above on these $2/\eta$ points in time $O(\eta^{-2})$ and derive $\tilde{f}(x)$. However, exact metric nearest neighbor search is potentially expensive, and so we cannot find these points efficiently. We instead find for each j a point $x' \in S$ with label $\tilde{f}(x') = j\eta/2$ that is a

$(1 + \frac{\eta}{2})$-approximate nearest neighbor of x among points with this label. (This can be done by presorting the points of S into $2/\eta$ buckets based on their $\tilde{f}$ label, and once x is received, running on each bucket a $(1 + \frac{\eta}{2})$-approximate nearest neighbor search algorithm due to [CG06] that takes $\eta^{-O(\mathrm{ddim}(\mathcal{X}))} \log n$ time.) We then run the brute force algorithm on these $2/\eta$ points in time $O(\eta^{-2})$. The nearest neighbor search achieves approximation factor $1 + \frac{\eta}{2}$, implying a similar multiplicative approximation to L, and thus also to $|y - f(x')| \leq 1$, which means at most $\eta/2$ additive error in the value y. We conclude that the algorithm's output solves the Lipschitz extension problem with additive η. □

References

[ABCH97] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

[BBL05] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of recent advances. *ESAIM Probab. Statist.*, 9:323–375, 2005.

[BFOS84] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984.

[BKL06] A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor. In *23rd international conference on Machine learning*, pages 97–104. ACM, 2006.

[CG06] R. Cole and L.-A. Gottlieb. Searching dynamic point sets in spaces with bounded doubling dimension. In *38th annual ACM symposium on Theory of computing*, pages 574–583, 2006.

[Cla99] K. L. Clarkson. Nearest neighbor queries in metric spaces. *Discrete Comput. Geom.*, 22(1):63–93, 1999.

[Cla06] K. Clarkson. Nearest-neighbor searching and metric space dimensions. In G. Shakhnarovich, T. Darrell, and P. Indyk, editors, *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*, pages 15–59. MIT Press, 2006.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1996.

[GK10] L.-A. Gottlieb and R. Krauthgamer. Proximity algorithms for nearly-doubling spaces. In *APPROX-RANDOM*, pages 192–204, 2010.

[GKK10] L.-A. Gottlieb, L. Kontorovich, and R. Krauthgamer. Efficient classification for metric data. In *COLT*, pages 433–440, 2010.

[GKK11] L.-A. Gottlieb, A. Kontorovich, and R. Krauthgamer. Efficient regression in metric spaces via approximate Lipschitz extension, 2011. `http://arxiv.org/abs/1111.4470`.

[GKK13] L.-A. Gottlieb, A. Kontorovich, and R. Krauthgamer. Adaptive metric dimensionality reduction, 2013. `http://arxiv.org/abs/1302.2752`.

[GKKW02] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[GKL03] A. Gupta, R. Krauthgamer, and J. R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *FOCS*, pages 534–543, 2003.

[HM06] S. Har-Peled and M. Mendel. Fast construction of nets in low-dimensional metrics and their applications. *SIAM Journal on Computing*, 35(5):1148–1184, 2006.

[KD11] S. Kpotufe and S. Dasgupta. A tree-based regressor that adapts to intrinsic dimension. *Journal of Computer and System Sciences, to appear*, 2011.

[KL04] R. Krauthgamer and J. R. Lee. Navigating nets: Simple algorithms for proximity search. In *15th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 791–801, January 2004.

[Kpo09] S. Kpotufe. Fast, smooth and adaptive regression in metric spaces. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1024–1032. 2009.

[KSW09] J. Kleinberg, A. Slivkins, and T. Wexler. Triangulation and embedding using small sets of beacons. *J. ACM*, 56:32:1–32:37, September 2009.

[LW08] J. Lafferty and L. Wasserman. Rodeo: Sparse, greedy nonparametric regression. *Ann. Stat.*, 36(1):28–63, 2008.

[LZ95] G. Lugosi and K. Zeger. Nonparametric estimation via empirical risk minimization. *IEEE Transactions on Information Theory*, 41(3):677–687, 1995.

[McS34] E. J. McShane. Extension of range of functions. *Bull. Amer. Math. Soc.*, 40(12):837–842, 1934.

[MH04] H. Q. Minh and T. Hofmann. Learning over compact metric spaces. In *COLT*, pages 239–254, 2004.

[Nad89] È. A. Nadaraya. *Nonparametric estimation of probability densities and regression curves*, volume 20 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht, 1989. Translated from the Russian by Samuel Kotz.

[Ney06] T. Neylon. *Sparse solutions for linear prediction problems*. PhD thesis, New York University, 2006.

[Pol84] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[SBWA98] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

[Tsy04] A. B. Tsybakov. *Introduction à l'estimation non-paramétrique*, volume 41 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2004.

[Vap95] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, New York, 1995.

[vLB04] U. von Luxburg and O. Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004.

[Was06] L. Wasserman. *All of nonparametric statistics*. Springer Texts in Statistics. Springer, New York, 2006.

[Whi34] H. Whitney. Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society*, 36(1):pp. 63–89, 1934.

[You01] N. E. Young. Sequential and parallel algorithms for mixed packing and covering. In *In 42nd Annual IEEE Symposium on Foundations of Computer Science*, pages 538–546, 2001.